UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Escola d'Enginyeria de Barcelona Est

TREBALL FI DE GRAU

**Grau en Enginyeria Biomèdica**

# ANÀLISI DE DADES A TWITTER SOBRE LA DIFUSIÓ D'INFORMACIÓ SOBRE EL DIÒXID DE CLOR (MMS/CDS) COM A TRACTAMENT PER A LA SALUT DE LES PERSONES

## Memòria i Annexos


**Autor:** Ariadna Costas Mañero
**Director:** Joan Martinez Sanchez


**Convocatòria:** Juny 2021

# Resum

L'objectiu principal del treball és implementar i avaluar diversos algoritmes de classificació de sentiment per a un cas concret de 'fake new' en salut, com és l'ús de MMS com a teràpia. En concret, s'ha optat per realitzar un anàlisi basat en el lèxic amb diccionaris, i un altre amb algoritmes d'aprenentatge supervisat.

Per al primer anàlisi, inicialment s'han utilitzat uns diccionaris estàndard, que posteriorment s'han modificat per fer més acurada la classificació. Els resultats mostren que, amb aquests canvis, s'augmenta el nombre de dades classificades (8,5%). També sembla que l'exactitud de les classificacions millora, però no es pot afirmar amb seguretat segons les dades de les que es disposa.

Els algoritmes d'aprenentatge supervisat que s'utilitzen són Decision Tree i Random Forest. Les dades d'exactitud són, respectivament, 0,922 i 0,996. S'apliquen mètodes de validació creuada per tal de verificar que els algoritmes estiguin correctament ajustats.

Les prediccions dels dos anàlisis mostren que els algoritmes d'aprenentatge supervisat obtenen una exactitud més elevada i que aconsegueixen classificar totes les dades, al contrari que el mètode basat en diccionaris. Però aquest últim resulta més útil a l'hora de realitzar classificacions sense etiquetar, una funció de la que els altres algoritmes no disposen.

# Resumen

El objetivo principal del trabajo es implementar y evaluar diversos algoritmos de clasificación de sentimiento para un caso concreto de 'fake new' en salud, como es el uso de MMS como terapia. En concreto, se ha optado por realizar un análisis basado en el léxico con diccionarios, y otro con algoritmos de aprendizaje supervisado.
Para el primer análisis, inicialmente se han utilizado unos diccionarios estándar, que posteriormente se han modificado para hacer más precisa la clasificación. Los resultados muestran que, con estos cambios, se aumenta el número de datos clasificados (8,5%). También parece que la exactitud de las clasificaciones mejora, pero no se puede afirmar con seguridad según los datos de los que se dispone.
Los algoritmos de aprendizaje supervisado que se utilizan son Decision Tree y Random Forest. Los datos de exactitud son, respectivamente, 0,922 y 0,996. Se aplican métodos de validación cruzada para verificar que los algoritmos estén correctamente ajustados.
Las predicciones de los dos análisis muestran que los algoritmos de aprendizaje supervisado obtienen una exactitud más elevada y que consiguen clasificar todos los datos, al contrario que el método basado en diccionarios. Pero este último resulta más útil a la hora de realizar clasificaciones sin etiquetar, una función de la que los otros algoritmos no disponen.

# Abstract

The main objective of this project is to implement and evaluate some sentiment classification algorithms for a specific case of fake new in health, such as the use of MMS as therapy. Specifically, a lexicon-based analysis with dictionaries and another with supervised learning algorithms are performed.
For the first analysis, standard dictionaries were initially used, which were later modified to make the classification more accurate. The results show that with these changes, the number of classified data increases (8.5%). It also seems that the accuracy of the classifications is improving, but it can not be stated with certainty according to the available data.
The supervised learning algorithms used are Decision Tree and Random Forest. Accuracy data are 0.922 and 0.996, respectively. Cross-validation methods are applied to verify that the algorithms are correctly adjusted.
The predictions in both analyzes show that supervised learning algorithms achieve higher accuracy and are able to classify all data, as opposed to the dictionary-based method. But the latter is more useful when performing unlabeled classifications, a function that other algorithms do not have.

# Índex

# 1. Introducció

Les 'fake news' en la salut són un problema important de desinformació. Les estadístiques mostren que han augmentat els últims anys i que moltes d'elles es transmeten principalment a través de les xarxes socials. Donat el gran volum de dades que es genera en aquestes xarxes, per poder comprendre la dinàmica de la 'fake new' i quina opinió té la gent al respecte calen aplicacions d'aprenentatge automàtic per classificar els missatges.

L'anàlisi de sentiment pot ser d'ajuda a l'hora d'entendre la opinió general sobre un determinat tema. Posteriorment, amb una segmentació dels resultats per usuaris, es pot conèixer quin és el perfil dels que creuen i promouen 'fake news', per exemple, per desenvolupar estratègies que ajudin a combatre la desinformació.

Per això aquest treball busca realitzar un anàlisi de sentiment amb diverses tècniques, i poder saber el seu rendiment i viabilitat d'aplicació.

## 1.1. Objectius del treball

- Implementació d'algoritmes basats en el lèxic per a l'anàlisi de sentiment en un cas relacionat amb les fake news en salut, en concret sobre el MMS
- Implementació d'algoritmes classificadors d'aprenentatge supervisat per a l'anàlisi de sentiment en un cas relacionat amb les fake news en salut, en concret sobre el MMS
- Avaluació dels rendiment dels algoritmes anteriors

## 1.2. Abast del treball

Amb aquest treball es pretén realitzar un anàlisi de sentiment des de dues perspectives diferents: amb algoritmes basats en el lèxic i amb aprenentatge supervisat. Per al primer, s'utilitzarà un mètode basat en diccionaris i per al segon es provaran diversos algoritmes per comprovar quins funcionen i s'ajusten millor a les dades de les que disposem.

L'anàlisi serà binari, és a dir, pretendrà classificar les dades en dos possibles sentiments: positiu o negatiu (envers l'ús del MMS com a teràpia). Les dades per les quals no es tingui clar a quina categoria pertanyen, ja sigui perquè no expressin cap opinió, siguin confuses o el text sigui massa curt, es descartaran.

L'objecte a analitzar serà el text sencer dels missatges de Twitter. Com que es tracta de missatges curts, es realitzarà una simplificació i no es considerarà la possibilitat que un missatge pugui contenir més d'un sentiment. Tots els missatges que s'analitzin estaran escrits en el mateix idioma i seran de la mateixa temàtica.

# 2. Marc teòric

## 2.1. 'Fake news' de salut

Les 'fake news' o notícies falses són aquelles notícies que contenen informació enganyosa o inexacta i contribueixen a la desinformació [1]. Segons un estudi realitzat l'any 2019, que es basa en una enquesta a més de 300 professionals de salut indica que bona part dels metges (62%) ha detectat un increment en els últims anys de les fake news de salut; d'aquests, un 77% creuen que és una conseqüència de l'augment de l'ús de les xarxes socials [2]. Recolza aquesta idea un article referent a les fake news sobre COVID-19, que analitza 1225 notícies falses, va concloure que la meitat d'aquestes es difonien principalment a través de les xarxes socials [1].

Les causes principals d'aquest fet són, per una banda, la falta de validació de les fonts que les promouen [2] i, de l'altra, la facilitat d'accés a internet i a les xarxes.

En relació als temes, s'estima que dos de cada tres fan referència a algun tipus de pseudoteràpia [2]. Les conseqüències d'això van des d'un increment en les consultes a professionals directament relacionades amb el tema [2], fins a casos d'efectes adversos per consum d'aquestes substàncies [3].

## 2.2. El diòxid de clor (MMS)

El MMS (Solució Mineral Miraculosa, per les sigles en anglès) és una solució al 28% de clorit sòdic, que quan es fa reaccionar amb certs àcids dona lloc al diòxid de clor. També s'anomena CDS (sigles de Solució de Diòxid de Clor), que és la mateixa substància però amb una altra presentació. Aquests dos compostos es promouen com a tractament per a malalties i trastorns com ara: la malària o paludisme, la COVID-19, l'autisme, el càncer o la diabetis, entre molts d'altres [4].

El seu primer promotor va ser l'enginyer Jim Humble, que l'any 2006 va publicar el llibre *The Miracle Mineral Solution of the 21st Century* [4]. Des de llavors, la seva popularitat ha anat creixent i s'han creat organitzacions que defensen, venen i promouen el seu ús, com ara l'Església del Gènesis II, Dolça Revolució, o Nova Medicina Germànica. Totes elles compten amb perfils de Facebook, Twitter o grups de Telegram per donar-se a conèixer.

L'any 2020, amb la pandèmia de la COVID-19, va augmentar el nombre de mencions a les xarxes socials (més endavant es pot veure amb les dades de Twitter de l'apartat **4.1 Obtenció de dades**), ja que molt aviat es va promocionar com la cura del virus. Apareix COMUSAV (Coalición Mundial Salud y Vida), una

organització de metges de Sud-Amèrica i Espanya que ofereix seminaris i protocols sobre com administrar la substància en casos de COVID-19.

Hi ha registrat dos assajos clínics que utilitzin el diòxid de clor com a medicament per ingestió [5] [6], un en fase II i l'altre amb resultats positius publicats [7]. No obstant això, els dos es centren en la malaltia per COVID-19, no s'han trobat assajos clínics amb d'altres patologies.

La majoria d'agències reguladores de medicaments no n'ha aprovat el seu ús com a medicament per a humans. És més, l'Agència Espanyola de Medicament i Productes Sanitaris [8] i la FDA (U.S. Food and Drug Administration) [9], entre d'altres organismes, porten emetent des de 2010 diversos avisos alertant dels perills del seu consum. Ambdós informen que és un medicament no autoritzat, i que pot provocar reaccions adverses greus tals com insuficiència hepàtica, intoxicacions o deshidratació [9].

Amb tot això, podem concloure que la seva promoció per als usos esmentats anteriorment conforma una 'fake new' en l'àmbit de les pseudoteràpies, ja que és una substància no aprovada com a medicament (i en alguns lloc com a Espanya està prohibit) i que té una manca d'evidència científica per a la majoria d'usos pels que es promou.

## 2.3. Twitter

Twitter és una xarxa social amb 353 milions d'usuaris, que cada dia genera un volum de dades aproximat de 656 milions de tweets [10]. Els missatges que s'hi publiquen són de 280 caràcters com a màxim, i qualsevol persona pot accedir-hi amb un compte, veure i publicar missatges.

El primer motiu pel qual s'ha utilitzat aquesta xarxa és que els seus missatges són públics. Això és imprescindible per poder obtenir les dades que necessitem per fer l'anàlisi. Però, a més, l'altre avantatge que presenta és que els missatges que s'hi publiquen són curts, cosa que ens facilitarà l'anàlisi de sentiment, ja que els textos llargs normalment són més complexos de classificar per a un algoritme.

Per últim, Twitter compta amb una API de desenvolupador que permet descarregar dades directament de la plataforma (requereix d'algun altre software, com es comenta més endavant). És a dir, està preparada per aquest tipus d'aplicacions i ofereix facilitats per a realitzar-les.

### 2.3.1. Marc legal

Per utilitzar Twitter API per a qualsevol finalitat, és necessari acceptar els seus termes, condicions i restriccions d'ús. Les principals restriccions que són d'aplicació al treball tenen a veure amb la informació personal dels usuaris a la qual es té accés (nom, nom d'usuari i identificador). No està permès utilitzar aquestes dades per determinar informació sensible d'un usuari concret tal com: afiliacions polítiques, estat de salut, creences religioses o situació financera, entre d'altres. Sí que es permet analitzar aquesta informació de manera agregada, sense correlacionar-ho amb cap dada de l'usuari [11].

En el treball només està previst realitzar un anàlisi de sentiment sobre una 'fake new' amb el text dels tweets, sense relacionar el sentiment del missatge amb les dades personals.

## 2.4. Mineria de text i anàlisi de sentiment

L'anàlisi de text o text mining té com a característica que opera amb el llenguatge natural, que són dades no estructurades. Com indica el seu nom, aquest tipus de dades no tenen una estructura interna definida, com sí que és el cas de les dades numèriques o les categòriques.

L'anàlisi de sentiment és un tipus d'anàlisi de text, que té com a objectiu definir eines automàtiques que extreguin informació subjectiva de textos en llenguatge natural, com són les opinions i els sentiments, per crear coneixement estructurat [12].

Aquest consta de múltiples característiques, algunes estan explicades a continuació:

- **Categories de sentiment:** El sentiment d'un text, en primer lloc, es pot classificar entre objectiu o subjectiu. Dins d'aquest últim cas, que normalment està format per opinios, es pot classificar la polaritat del sentiment, que pot ser positiu, negatiu o neutral. També es poden afegir altres categories depenent les dades i de l'ús que se'n vulgui fer, com per exemple una valoració de l'1 al 5 , com es fa en moltes ressenyes [12].
- **Nivell d'anàlisi:** És la definició de l'objecte a analitzar. Aquest pot ser el missatge sencer (s'atribueix un únic sentiment al mateix), les frases (en cas que el missatge en tingui vàries, es pot analitzar cada una d'elles per separat i obtenir una categorització del missatge amb més matisos) o les entitats (o paraules, divideixen el text en fragments amb el mateix sentiment i estableixen quina és la paraula clau de cada un). [12]
- **Tipus d'opinió:** Aquesta pot venir expressada de manera directa, indirecta o comparativa entre dos termes. [12]

En funció de les característiques de les dades i el tipus de resultats que es vulguin obtenir, es poden implementar diferents mètodes per analitzar el text. Alguns dels més utilitzats en l'anàlisi de sentiment es troben en la Figura 1:

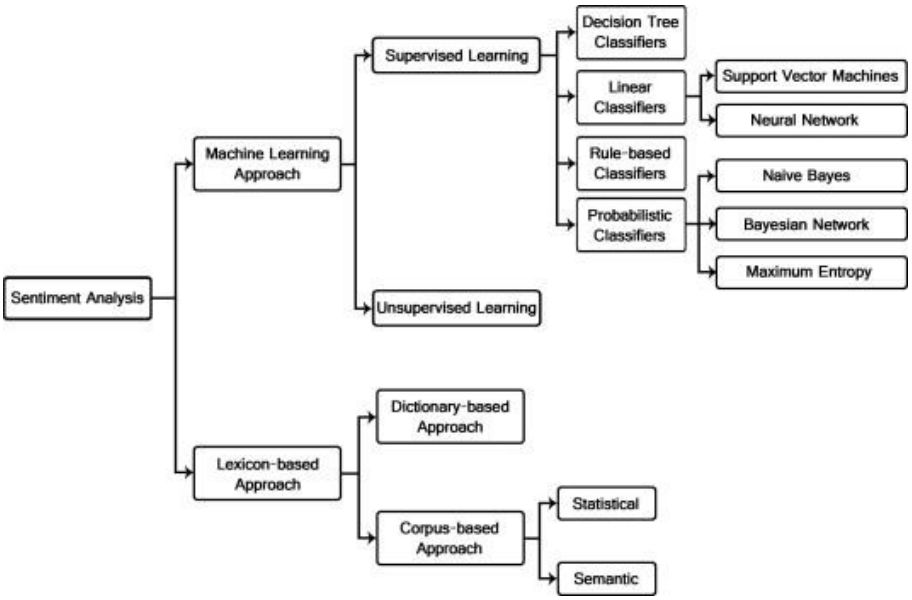


*Figura 1: Mètodes utilitzats per l'anàlisi de sentiment. (Font: Medhat et. al. [13])*

Com podem veure, els algoritmes més utilitzats són els basats en el lèxic i els d'aprenentatge supervisat.

# 3. Mètodes

## 3.1. Procediment

El procediment del treball es dividirà en tres parts principals: l'extracció de dades, l'anàlisi de sentiment (incloent el preprocessament de les dades) i la posterior validació del mateix.

La primera consistirà en obtenir un nombre de dades suficient i amb prou variables com per realitzar un anàlisi de sentiment. L'extracció de dades es portarà a terme amb les eines Twitter API i Python. L'objectiu serà extreure'n un gran volum (més de 10.000 dades) per tal d'assegurar el correcte funcionament de l'aplicació. En aquesta etapa també s'eliminaran aquelles dades que no compleixin un certs requisits o que estiguin repetides i es filtraran les variables que no siguin d'interès per al desenvolupament del treball.

Posteriorment, amb la plataforma KNIME es realitzaran dos tipus d'anàlisi de sentiment, un d'ells basat en diccionaris i l'altre en algoritmes d'aprenentatge supervisat. L'objectiu d'aquests serà etiquetar les dades segons siguin positives (és a dir, que estiguin a favor del MMS o diòxid de clor) i negatives (que hi estiguin en contra).

El primer anàlisi tracta de determinar el sentiment d'un missatge (dada) a partir del lèxic del mateix, detectant quines paraules tenen una connotació positiva i quines negativa i fent un balanç per determinar-ne el sentiment general. Per realitzar-lo, només cal disposar dels textos que volem analitzar (no és necessària ni admet cap altra variable), però també requereix de llistes (diccionaris) predefinides que indiquin quines paraules són positives o negatives.

L'anàlisi amb algoritmes d'aprenentatge supervisat funciona de manera diferent; a partir d'un conjunt de dades que ja estan etiquetades amb el respectiu sentiment (o categoria), es crea un model que s'ajusti a aquestes dades, i que idealment hauria de predir correctament el sentiment d'un nou conjunt de dades. Donada la versatilitat d'aquests models és possible incloure diverses variables a part del text. Però presenta l'inconvenient que necessita de dades anteriorment etiquetades per tal d'entrenar el model.

Sabent que disposarem d'una gran quantitat de dades i que aquestes no estaran etiquetades, i coneixent que hauríem de disposar de com a mínim un 70% de les dades etiquetades perquè el model funcioni de manera òptima, podem concloure que és poc viable realitzar-ho manualment. Per aquest motiu aplicarem aquest anàlisi amb les dades que obtindrem de l'anàlisi amb diccionaris.

Per últim, es validaran els dos models i es discutiran els resultats obtinguts.

## 3.2. Eines

### 3.2.1. KNIME

KNIME Analytics Platform és un software de ciència de dades. Permet llegir i exportar arxius de múltiples formats, filtrar i modificar dades, realitzar càlculs estadístics, a més, conté diversos algoritmes de machine learning [14]. Els seus programes s'anomenen workflows i la seva estructura es basa en unitats anomenades nodes, cada un d'ells realitza una acció específica i la majoria permeten canviar la seva configuració.

L'aplicació principal del treball es desenvoluparà en aquesta plataforma, pels següents motius:

- Domini de l'eina, encara que no en mineria de text
- És un software lliure
- És àmpliament utilitzat, per aquest motiu hi ha una gran quantitat d'exemples, llibres, vídeos...
- Compta amb KNIME FORUM, amb experts de la comunitat que poden resoldre dubtes sobre l'ús del programa
- Compta també amb KNIME HUB, amb workflows d'accés obert
- Permet integrar la API de Twitter
- Permet integrar altres llenguatges de programació si és necessari, com Python o R.

### 3.2.2. Twitter API

Twitter API és una aplicació de la xarxa social Twitter que s'utilitza per analitzar dades, publicar i rebre missatges des de diferents llenguatges de programació.

Per accedir a un compte de desenvolupador de l'aplicació cal tenir un compte a Twitter, i respondre a un seguit de preguntes sobre l'ús que se'n vol fer. Entre elles s'inclou: si es pretén analitzar informació de Twitter, quines dades s'agafaran i amb quines tècniques es farà l'anàlisi, si es pretén interactuar amb comptes d'altres usuaris i com es vol fer, i si el contingut proporcionat es publicarà fora de Twitter.

El compte estàndard de desenvolupador permet, entre d'altres coses, obtenir tweets dels últims 7 dies, amb un màxim de 100 missatges per consulta i 500 consultes al mes. A més, no es poden realitzar més de 75 consultes en un període de 15 minuts.

Si es volen obtenir dades anteriors als últims 7 dies, cal tenir un compte Premium, hi ha dues opcions: una subscripció que permet obtenir tweets dels últims 30 dies i una altra que dona accés a tota la base

de dades. Ambdós permeten 500 missatges per consulta, i un màxim de 60 consultes per minut. El nombre de consultes al mes dependrà de la subscripció, n'hi ha des de 100 fins a 10.000.

S'hi pot accedir des de diferents llenguatges de programació, per tal d'identificar-se calen cinc credencials: dues claus (API Key i API Secret, serveixen per identificar l'usuari), i tres "tokens" (Access token, Access token secret i Bearer token) que són un altre tipus de credencials que identifiquen l'aplicació. Totes aquestes credencials són generades per Twitter i es poden regenerar tants cops com es vulgui, però no es poden modificar.

### 3.2.3. Python

Python és un llenguatge de programació d'ús lliure, que conté diverses llibreries i paquets per enllaçar amb d'altres programes (entre ells Twitter API).

Tot i que la idea inicial era utilitzar KNIME com a únic programa de desenvolupament de l'aplicació, algunes limitacions d'aquest van provocar que l'obtenció de dades es portés a terme amb Python. La principal limitació era que els nodes de KNIME que connecten amb Twitter API estan configurats per obtenir dades únicament dels últims 7 dies. Per aquest motiu, quan es va decidir optar per cercar a l'arxiu complet de Twitter es va buscar una altra eina per poder extreure les dades necessàries.

# 4. Metodologia

## 4.1. Obtenció de dades

Per obtenir dades de Twitter, KNIME té un node mitjançant el qual es pot accedir a la API de Twitter (*Twitter API Connector*) i un altre (*Twitter Search*) amb el que es poden realitzar cerques de tweets amb unes característiques comunes. No obstant això, aquests nodes no permeten fer cerques amb la versió Premium de Twitter, amb la qual cosa s'ha optat per utilitzar un altre eina (Python) per obtenir aquestes dades. Aquestes s'obtenen per defecte en format JSON Lines (jsonl), que més endavant es processa amb KNIME.

Inicialment, es va optar per utilitzar el compte estàndard, i anar realitzant consultes un cop per setmana. D'aquesta manera, només es podrien obtenir dades del període en que s'estava desenvolupant el treball. Les dues primeres setmanes es van obtenir menys de 100 tweets cada una, tots els que hi havia referents al tema (50 missatges del 2 al 8 de març, i 65 del 9 al 15 de març). Suposant una mitjana de 50 tweets per setmana i que es podrien prendre dades fins a finals de maig (13 setmanes), obtenim una estimació de 650 missatges en total. Amb aquest volum de dades, és possible que no es puguin obtenir uns resultats satisfactoris de l'anàlisi de sentiment, a més que no hi ha cap garantia que es mantingui la xifra de 50 missatges a la setmana. Per evitar arribar al final del període de realització del treball sense dades suficients, es va decidir ampliar la subscripció a Premium. La que es va escollir permet accedir a tota la base de dades, i té un límit de 100 consultes al mes, per tant, es poden obtenir fins a 50.000 dades (perquè cada consulta té un límit de 500 dades).

En qualsevol dels casos, la informació que Twitter subministra inclou les següents característiques (variables) de cada tweet:

- Contingut
- Identificador
- Dia i hora de creació
- Si es una resposta o retweet, i en cas afirmatiu l'identificador del tweet original
- Nom d'usuari i descripció del compte autor del tweet, la seva localització i la data de creació
- Idioma
- Altra informació de menys interès per a la aplicació que es vol desenvolupar.

Per obtenir les dades cal especificar una consulta (com ara quines paraules volem que contingui el missatge o de quin usuari provenen els missatges) i també les dates entre les quals es vol fer la consulta. A part, per desar la base de dades cal especificar el nombre màxim de missatges que es volen obtenir.

En relació a la consulta, com que les sigles MMS fan referència a molts termes, per tal de comprovar que només s'obtinguin tweets de la temàtica desitjada s'introdueix al buscador que, a més que el tweet contingui MMS, també hauria de contenir algun dels següents termes: CDS, dioxidodecloro o chlorinedioxide. És interessant notar que aquesta cerca no ens proporcionarà tots els missatges que s'hagin escrit sobre el tema que volem estudiar, ja que pot haver-hi usuaris que únicament mencionin el terme "MMS" o "CDS" en el seu missatge sense fer referència als altres paraules relacionades. Tot i això, s'ha considerat que és millor perdre dades vàlides que no pas obtenir dades no relacionades amb el tema i haver-les de filtrar posteriorment.

Especificar les dates de la consulta i el nombre màxim de tweets és un procés més complex, per diverses raons. La principal és la limitació de consultes, que com s'ha comentat abans és de 100 al mes, i cada consulta pot contenir un màxim de 500 tweets. En cas que hi hagi menys de 500 missatges que compleixin la condició especificada a la consulta, aquesta es comptabilitzarà igualment però s'obtindran menys dades. Per aquest motiu és important tenir una estimació de quants missatges es podran obtenir en un període de temps, per no malbaratar consultes. Això es calcula mitjançant el mètode de prova i error.

Cada una de les consultes realitzades s'ha desat en un arxiu, obtenint un total de 29 documents. Tot i que hagués estat possible desar totes les dades en un sol document, i probablement hagués estat més còmode a l'hora de processar-lo, el programa funciona més ràpid creant arxius nous per a cada consulta que obrint i modificant els que ja s'han desat.

En total s'han obtingut 19804 dades d'entre gener de 2011 i març de 2021, la seva distribució al llarg del temps es pot trobar en la taula següent:

| **Període** | **Nombre de tweets** |
|---|---|
| **Gener 2021 – Març 2021** | 2287 |
| **Octubre 2020 – Desembre 2020** | 1908 |
| **Juliol 2020 – Setembre 2020** | 8500 |
| **Abril 2020 – Juny 2020** | 4690 |
| **Gener 2020 – Març 2020** | 412 |
| **Gener 2019 – Desembre 2019** | 370 |
| **Gener 2018 – Desembre 2018** | 288 |

| | |
|---|---|
| **Gener 2017 – Desembre 2017** | 258 |
| **Gener 2016 – Desembre 2016** | 185 |
| **Gener 2015 – Desembre 2015** | 237 |
| **Gener 2014 – Desembre 2014** | 292 |
| **Gener 2013 – Desembre 2013** | 167 |
| **Gener 2012 – Desembre 2012** | 151 |
| **Gener 2011 – Desembre 2011** | 59 |
| **Total** | 19804 |

No es van cercar dades anteriors a 2011, ja que eren molt escasses.

Finalment, es mostra una de les dades obtingudes en format JSON Lines, per mostrar la informació que s'ha obtingut de cada missatge (s'han eliminat els camps de l'usuari: id, id_str, name i screen_name per complir amb la protecció de dades de Twitter):


```
{"created_at": "Mon Mar 29 22:37:08 +0000 2021", "id":
1376664749358911492, "id_str": "1376664749358911492", "text": "CDS y
MMS\nOxidante y Oxigenante\n\nElimina bacterias, virus, hongos y
par\u00e1sitos de tu organismo.\nAyuda en el control\u2026
https://t.co/CNaXeysRs4", "display_text_range": [0, 140], "source": "<a
href=\"https://postcron.com\" rel=\"nofollow\">Postcron App</a>",
"truncated": true, "in_reply_to_status_id": null,
"in_reply_to_status_id_str": null, "in_reply_to_user_id": null,
"in_reply_to_user_id_str": null, "in_reply_to_screen_name": null,
"user": {"id": XXX, "id_str": "XXX", "name": "XXX", "screen_name":
"XXX", "location": null, "url": "http://www.periodicoaqui.com",
"description": "Ya no imprimimos Gacetas, ahora somos Revista Digital
con m\u00e1s de 1500 ediciones publicadas, somos su mejor opci\u00f3n
publicitaria desde hace m\u00e1s de 31 a\u00f1os.", "translator_type":
"none", "protected": false, "verified": false, "followers_count": 6309,
"friends_count": 5894, "listed_count": 17, "favourites_count": 757,
"statuses_count": 147465, "created_at": "Fri May 28 04:36:06 +0000
2010", "utc_offset": null, "time_zone": null, "geo_enabled": true,
"lang": null, "contributors_enabled": false, "is_translator": false,
"profile_background_color": "C0DEED", "profile_background_image_url":
"http://abs.twimg.com/images/themes/theme1/bg.png",
"profile_background_image_url_https":
"https://abs.twimg.com/images/themes/theme1/bg.png",
"profile_background_tile": true, "profile_link_color": "0084B4",
"profile_sidebar_border_color": "C0DEED", "profile_sidebar_fill_color":
"DDEEF6", "profile_text_color": "333333",
"profile_use_background_image": true, "profile_image_url":
"http://pbs.twimg.com/profile_images/1198241356/ipaper_normal.jpg",
```

```
"profile_image_url_https":
"https://pbs.twimg.com/profile_images/1198241356/ipaper_normal.jpg",
"profile_banner_url":
"https://pbs.twimg.com/profile_banners/149001262/1558746287",
"default_profile": false, "default_profile_image": false, "following":
null, "follow_request_sent": null, "notifications": null,
"withheld_in_countries": []}, "geo": null, "coordinates": null,
"place": null, "contributors": null, "is_quote_status": false,
"extended_tweet": {"full_text": "CDS y MMS\nOxidante y
Oxigenante\n\nElimina bacterias, virus, hongos y par\u00e1sitos de tu
organismo.\nAyuda en el control de diversas
patolog\u00edas.\n\n\u00a1Ll\u00e1manos! 556967-1195 / 558369-
7497\nEnv\u00edos a toda la Rep\u00fablica\nSalud posible
https://t.co/DSGgIJoEHJ", "display_text_range": [0, 217], "entities":
{"hashtags": [], "urls": [], "user_mentions": [], "symbols": [],
"media": [{"id": 1376664747291127809, "id_str": "1376664747291127809",
"indices": [218, 241], "media_url":
"http://pbs.twimg.com/media/ExrlSzUXMAEmMt_.jpg", "media_url_https":
"https://pbs.twimg.com/media/ExrlSzUXMAEmMt_.jpg", "url":
"https://t.co/DSGgIJoEHJ", "display_url": "pic.twitter.com/DSGgIJoEHJ",
"expanded_url":
"https://twitter.com/periodicoAQUI/status/1376664749358911492/photo/1",
"type": "photo", "sizes": {"thumb": {"w": 150, "h": 150, "resize":
"crop"}, "small": {"w": 680, "h": 443, "resize": "fit"}, "large": {"w":
1104, "h": 720, "resize": "fit"}, "medium": {"w": 1104, "h": 720,
"resize": "fit"}}}]}, "extended_entities": {"media": [{"id":
1376664747291127809, "id_str": "1376664747291127809", "indices": [218,
241], "media_url": "http://pbs.twimg.com/media/ExrlSzUXMAEmMt_.jpg",
"media_url_https": "https://pbs.twimg.com/media/ExrlSzUXMAEmMt_.jpg",
"url": "https://t.co/DSGgIJoEHJ", "display_url":
"pic.twitter.com/DSGgIJoEHJ", "expanded_url":
"https://twitter.com/periodicoAQUI/status/1376664749358911492/photo/1",
"type": "photo", "sizes": {"thumb": {"w": 150, "h": 150, "resize":
"crop"}, "small": {"w": 680, "h": 443, "resize": "fit"}, "large": {"w":
1104, "h": 720, "resize": "fit"}, "medium": {"w": 1104, "h": 720,
"resize": "fit"}}}]}}, "quote_count": 0, "reply_count": 0,
"retweet_count": 1, "favorite_count": 0, "entities": {"hashtags": [],
"urls": [{"url": "https://t.co/CNaXeysRs4", "expanded_url":
"https://twitter.com/i/web/status/1376664749358911492", "display_url":
"twitter.com/i/web/status/1\u2026", "indices": [116, 139]}],
"user_mentions": [], "symbols": []}, "favorited": false, "retweeted":
false, "possibly_sensitive": false, "filter_level": "low", "lang":
"es", "matching_rules": [{"tag": null}]}
```


El codi utilitzat per a l'extracció de dades es pot trobar a l'Annex A i es va extreure íntegrament de la referència [15].

## 4.2. Construcció de la base de dades

Un cop obtinguts tots els arxius de dades, es construeix la base de dades en un de sol. El procediment en aquesta part és molt senzill; un cop s'han obtingut tots els documents, es llegeixen amb KNIME i es

concatenen per tal de formar una sola base de dades. Finalment es desaran amb el format .table, un format específic amb el que treballa KNIME.

La lectura dels documents es realitza amb el node *File Reader*. És important destacar que, tot i que en el fitxer original les dades no estaven numerades, al llegir-les se'ls assigna un número de fila que serveix com a identificador (a la primera dada se la identifica com a Row0, a la segona com a Row1 i així successivament).

Seguidament utilitzem el node *Concatenate*, que modificarem perquè pugui admetre 29 fitxers d'entrada (originalment el node n'admet 2). També permet la configuració d'altres paràmetres, el primer es refereix a què fer en cas que els identificadors de les files dues taules diferents siguin iguals. Hi ha l'opció d'eliminar-los, de mantenir-los o de no concatenar les taules (enviar un missatge d'error). Degut que s'han llegit els 29 fitxers de manera independent, hi ha diversos identificadors repetits i com que no ens interessa perdre les dades, escollim la segona opció. Això farà que, a cada columna que tingui un identificador repetit, se li afegeixi _dup a aquest. Addicionalment, també es pot configurar si es desitja que la taula resultant sigui la unió o la intersecció de les taules, escollim la primera opció pel motiu exposat anteriorment.

Per acabar amb aquesta part, desem en una taula els resultats, amb el node *Table Writer*. El motiu principal pel qual es desen els resultats en taules és poder anar fent còpies de seguretat de les dades obtingudes en diferents passos intermedis.

## 4.3. Selecció de les variables

Les dades que contenen més informació consten de 3985 variables. Però, per a fer l'anàlisi no es tindran totes en compte, per diversos motius:

- Algunes variables no estan especificades per totes les dades.
- Hi ha variables que no semblen rellevants a l'hora de fer l'anàlisi. En serien un exemple la data i hora de creació del compte o bé el seu color.
- La idea del treball és que sigui reproduïble, que el programa es pugui implementar amb molta menys informació, per exemple amb dades extretes amb el compte bàsic de Twitter API o amb dades d'altres xarxes socials. Per això s'intentarà utilitzar el mínim de variables possible.
- En general, és preferible utilitzar models amb poques variables, pel cost computacional i per evitar sobreajustaments.

Per tots aquests motius, la variable seleccionada és el text del tweet (columna "text"). És la única que és imprescindible per l'anàlisi de sentiment, i a més, es pot dividir en vàries variables, una per paraula. Es tracta d'una variable de tipus String.

Per filtrar la columna, s'utilitza el node *Column filter*. La seva configuració és molt senzilla, només cal seleccionar quines variables o columnes volem mantenir.

A part de la variable esmentada anteriorment, també es filtren algunes columnes que, si bé no s'utilitzaran per l'anàlisi de sentiment, seran útils en etapes posteriors a la selecció de variables. El primer és l'autor de tweet (tipus String, anomenat "user/screen name"), que ens servirà per agrupar-los més endavant. També se selecciona la columna de l'idioma ("lang"), que es farà servir per separar els missatges per aquest camp, ja que no es pot realitzar un anàlisi de sentiment que inclogui més d'una llengua.

Un cop tenim la taula amb totes les dades i les tres variables, es realitza un filtrat per idiomes. La seva finalitat és escollir un idioma concret per a realitzar l'anàlisi, que serà el que tingui més dades. Per fer-ho, s'aplica el node *Row Filter*, que permet filtrar dades que tinguin alguna característica comuna.

La taula resultant conté 16003 dades corresponents als tweets escrits en castellà, amb tres variables cadascuna (cap variable té un valor nul per a cap dada).

## 4.4. Anàlisi de sentiment: Diccionaris

### 4.4.1. Preprocessament

El preprocessament es defineix com el conjunt de tasques que transformen les dades en brut en un conjunt de dades útil i correcte per al processament. Aquest procés pot incloure suprimir informació irrellevant o redundant, ajuntar diverses bases de dades, imputar valors que falten i/o adaptar el format de les dades segons la tècnica amb que s'hagi de processar posteriorment [16].

Els nodes de KNIME dedicats al preprocessament de text únicament funcionen amb un tipus de variable anomenada Document. Per aquest motiu caldrà convertir la columna del text, això es pot fer fàcilment amb un node anomenat *Strings To Document*.

El format Document permet agrupar un conjunt de variables en una de sola, de manera que es pot emmagatzemar el text, l'autor i el tema del missatge (entre d'altres) en una columna. En el nostre cas, només desem el text. A part, el node *Strings To Document* també separa les paraules i les converteix en unitats independents. Aquest procés s'anomena 'Word Tokenization' [20], i KNIME permet escollir entre diversos mètodes per realitzar la separació. Triem que el criteri per dividir el text sigui els espais en blanc.

Separar les paraules és molt important ja que ens permet assignar una etiqueta a cada una d'elles de manera independent, cosa que no podríem fer si mantinguéssim el text en format String.

Un cop tenim les dades en un format apte per al preprocessament, com que per l'anàlisi de sentiment únicament ens interessa analitzar les paraules, s'eliminaran tots els nombres i signes de puntuació dels tweets. A més, es convertiran totes les lletres a minúscula, ja que això facilitarà a l'algoritme la detecció de paraules iguals. Per fer-ho, s'utilitzen els nodes *Number Filter*, *Punctuation Erasure* i *Case Converter*, cap dels tres necessita ser configurat.

Finalment, eliminem també aquelles paraules que no són determinants a l'hora de fer l'anàlisi de sentiment. En són exemples els connectors, que s'utilitzen en tots els missatges, tant positius com negatius. El que ens interessa són les paraules amb una càrrega explicativa, majoritàriament substantius, adjectius i verbs.

Això es porta a terme amb el node *Stop Word Filter*, que permet filtrar aquest tipus de paraules. Ofereix dues opcions; utilitzar un llistat de paraules ja configurades en el node (n'hi ha de diversos idiomes, entre ells el castellà) o utilitzar una taula pròpia. Seleccionarem la primera opció, ja que així ens evitem haver de realitzar una llista.

El document preprocessat s'emmagatzema en una nova columna anomenada 'Preprocessed Document'. Mantindrem també la columna amb el document original ('Document'), ja que aquesta és més fàcil de llegir (des d'un punt de vista humà) i ens resultarà útil en etapes posteriors.

| Document | Preprocessed Document |
|---|---|
| "CDS y MMSOxidante y OxigenanteElimina bacterias, virus, hongos y parásitos de tu organismo.Ayuda en el control... | "cds mmsoxidante oxigenanteelimina bacterias virus hongos parásitos tu organismoayuda con... |
| "@cleopatrarubia @DerEpr86 Estuvo año con MMS (2014) y mejoró mucho, voy a retomar el CDS, mil gracias!!!" | "cleopatrarubia dereprr86 año mms 2014 mejoró voy retomar cds mil gracias" |
| "@ldu2004 @AndresJ777 @TelevisionUpea Padeces titulitis. Has probado el MMS o el CDS? Yo sí.. Mi experiencia v... | "ldu2004 andresj777 televisionupea padeces titulitis has probado mms cds experiencia vale m... |
| "@de_dioxido Lo gracioso de esa web es que está eliminada la página que tenía guardada y que posteaba cada vez... | "dedioxido gracioso web eliminada página guardada posteaba a..." |
| "@El_Universal_Mx Que denuncien a los que atacan a los que defendemos y curamos con MMS y CDS, esos si son c... | "eluniversalmx denuncien atacan defendemos curamos mms cds criminale..." |
| "Menos mal que aún queda gente con la cabeza en su sitio allí. Gracias, @UMSABolivia. #MMS #CDS #EsLejía...." | "mal queda gente cabeza sitio allí gracias umsabolivia mms cds eslejía..." |
| "@Azariel8 @eluz_ra @alonsotoro Yo le puedo vender en la presentación que quiera. MMS o CDS. Escríbame al dire... | "azariel8 eluzra alonsotoro puedo vender presentación quiera mms cds escríbame directo cara... |
| "CDS y MMSOxidante y OxigenanteElimina bacterias, virus, hongos y parásitos de tu organismo.Ayuda en el control... | "cds mmsoxidante oxigenanteelimina bacterias virus hongos parásitos tu organismoayuda con... |

*Figura 2: Visualització dels documents inicial i els preprocessats*

### 4.4.2. Assignació d'etiquetes

Una vegada tenim el document preprocessat, el que farem és etiquetar algunes paraules del text amb algun dels dos sentiments. Aquestes etiquetes estaran associades a les paraules en qüestió i es mantindran encara que el document pateixi modificacions.

La manera com les etiquetarem serà automàtica, utilitzant dos diccionaris. Aquests consistiran en un llistat de paraules que habitualment tinguin una connotació negativa o positiva (segons el cas). Cada cop que una paraula d'un missatge coincideixi amb una que es troba al diccionari se li aplicarà l'etiqueta corresponent. Si classifiquem d'aquesta manera totes les paraules d'un tweet determinat i realitzem un balanç entre les paraules positives i negatives podrem concloure quin és el sentiment general del missatge.

Els diccionaris que utilitzem estan extrets de la referència [17] i els que s'han utilitzat es poden trobar a l'Annex B. En aquesta referència es poden trobar llistes de paraules positives i negatives separades per idiomes, entre 1500 i 3000 paraules per llistat.

El fet d'utilitzar un diccionari genèric presenta avantatges i inconvenients; d'una banda, com ja s'ha comentat, perquè permet la classificació automàtica de molts termes que habitualment són positius i negatius. Per altra banda, es dona també el cas que, una paraula que sol pertànyer a un dels dos sentiments, en el context que estem treballant no tingui aquesta connotació. Per exemple, la paraula "virus" es troba a la taula de les paraules negatives, però si llegim els missatges que contenen aquesta paraula, ens adonem ràpidament que en la majoria de casos tenen un significat positiu. Un parell d'exemples:

> *"La Ivermectina y el CDS o MMS te previenen de cualquier cepa o mutación del virus"*
>
> *" Mis experiencias también han sido magnificas, los virus se van en dos dias con el mms/CDS"*

Això també passa amb altres termes, de manera que les llistes s'hauran de modificar per tal que l'etiqueta de sentiment assignada sigui el més acurada possible.

El procediment que seguim per a l'assignació d'etiquetes és el següent: comencem llegint els dos diccionaris per separat amb dos nodes *File Reader*. Ambdós estan formats per una sola columna anomenada 'Col0' i, mentre que el de paraules negatives té 2720 files, el de positives consta de 1555. A continuació, per tal de modificar els diccionaris, afegim els nodes següents:

- *Table Creator*: Amb aquest node entrarem manualment paraules addicionals que vulguem incloure al diccionari. Són termes que no estan contemplats en els diccionaris originals, però que s'ha observat que majoritàriament tenen una connotació positiva o negativa. La

configuració té un format de taula, on es pot escriure directament. És important tenir en compte que s'ha de canviar el nom de la columna que hi ha per defecte, i posar el mateix que en el diccionari original ('Col0'). Si no, el següent node ho detectarà com a dues columnes diferents.

- *Concatenate:* Seguidament, creem una nova taula ajuntant el diccionari original amb les paraules anteriors. El resultat contindrà una sola columna i tantes files com paraules.
- *Rule-based Row Filter:* S'utilitza per eliminar aquelles paraules que estan incloses en els diccionaris originals però que no volem utilitzar perquè tenen un sentiment diferent de l'original en el nostre cas.

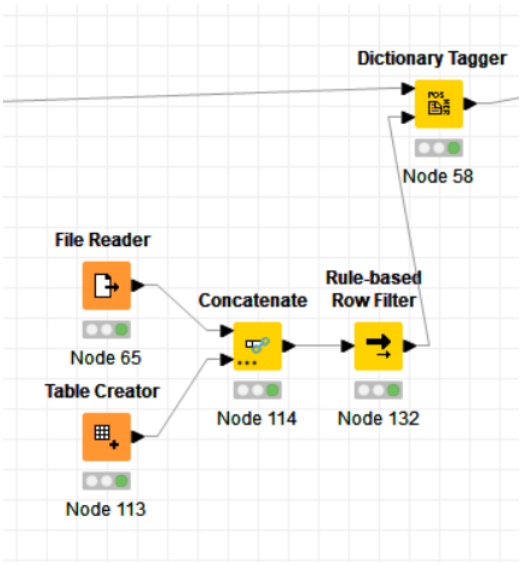


*Figura 3: Disposició dels nodes per a l'assignació i modificació del diccionari*

El procediment per determinar quines paraules afegir i quines treure de cada diccionari està explicat en l'apartat **4.4.3 Modificació dels diccionaris**.

Un cop tenim el diccionari definitiu, utilitzem el node *Dictionary Tagger*. Aquest té dues entrades: una per al diccionari i una altra per les dades a etiquetar (utilitzarem les dades preprocessades). Per configurar aquest node s'ha de posar el tipus d'etiqueta ('Sentiment'), el valor de l'etiqueta ('Positive' o 'Negative' segons el cas) i la columna en la que es troba el document ('Preprocessed Document').

El resultat és un document que conté algunes paraules etiquetades. Ara bé, presenta dos inconvenients: el primer és que no es pot visualitzar directament i el segon és que no s'etiqueta cada un dels missatges, si no les paraules. Per tant, cada missatge pot tenir més d'una etiqueta diferent, quan el que volem és assignar un únic valor per a cada dada.

Per solucionar el primer, el que farem serà separar els missatges (els documents preprocessats) en paraules (obtenir una taula que tingui una fila per paraula), i el següent pas serà afegir una columna a

cada paraula que es correspongui amb l'etiqueta assignada a aquesta. Precisament així transforma les dades el node *Bag of Words Creator*. Només cal configurar-lo amb la columna on tenim el text que volem separar i indicar quin nom volem que tingui la columna que afegirem, que serà 'Term'.

| Document | Preprocessed Document | Term |
|---|---|---|
| "CDS y MMSOxidante y OxigenanteElimina bacterias, virus, hongos y parásitos de tu organismo.Ayuda en el control... | "cds mmsoxidante oxigenanteelimina bacterias virus hongos parásitos tu organismoayuda control... | organismo[] |
| "CDS y MMSOxidante y OxigenanteElimina bacterias, virus, hongos y parásitos de tu organismo.Ayuda en el control... | "cds mmsoxidante oxigenanteelimina bacterias virus hongos parásitos tu organismoayuda control... | ayuda[POSITIVE(SENTIMENT)] |
| "CDS y MMSOxidante y OxigenanteElimina bacterias, virus, hongos y parásitos de tu organismo.Ayuda en el control... | "cds mmsoxidante oxigenanteelimina bacterias virus hongos parásitos tu organismoayuda control... | control...[] |
| "@cleopatrarubia @DerEpr86 Estuvo año con MMS (2014) y mejoró mucho, voy a retomar el CDS, mil gracias!!!" | "cleopatrarubia derepr86 año mms 2014 mejoró voy retomar cds mil gracias" | cleopatrarubia[] |
| "@cleopatrarubia @DerEpr86 Estuvo año con MMS (2014) y mejoró mucho, voy a retomar el CDS, mil gracias!!!" | "cleopatrarubia derepr86 año mms 2014 mejoró voy retomar cds mil gracias" | derepr86[] |
| "@cleopatrarubia @DerEpr86 Estuvo año con MMS (2014) y mejoró mucho, voy a retomar el CDS, mil gracias!!!" | "cleopatrarubia derepr86 año mms 2014 mejoró voy retomar cds mil gracias" | año[] |
| "@cleopatrarubia @DerEpr86 Estuvo año con MMS (2014) y mejoró mucho, voy a retomar el CDS, mil gracias!!!" | "cleopatrarubia derepr86 año mms 2014 mejoró voy retomar cds mil gracias" | mms[] |
| "@cleopatrarubia @DerEpr86 Estuvo año con MMS (2014) y mejoró mucho, voy a retomar el CDS, mil gracias!!!" | "cleopatrarubia derepr86 año mms 2014 mejoró voy retomar cds mil gracias" | 2014[] |
| "@cleopatrarubia @DerEpr86 Estuvo año con MMS (2014) y mejoró mucho, voy a retomar el CDS, mil gracias!!!" | "cleopatrarubia derepr86 año mms 2014 mejoró voy retomar cds mil gracias" | mejoró[] |
| "@cleopatrarubia @DerEpr86 Estuvo año con MMS (2014) y mejoró mucho, voy a retomar el CDS, mil gracias!!!" | "cleopatrarubia derepr86 año mms 2014 mejoró voy retomar cds mil gracias" | voy[] |

*Figura 4: Taula resultant del node Bag of Words Creator, amb el Document, el Document Preprocessat i cada un dels termes del text (alguns d'ells amb etiqueta de sentiment)*

A continuació, per tal de desar l'etiqueta de cada paraula en una columna a part utilitzem el node *Tags to String,* configurat perquè ens mostri l'etiqueta de sentiment per a cada paraula de la columna 'Term'. Alhora, es realitza també una taula de freqüències, de manera que es pot visualitzar quantes vegades apareix un terme en qüestió dins del document. La nova columna s'anomena "TF abs", i un fragment de la taula resultant es mostra a continuació:

| Preprocessed Document | Term | TF abs | SENTIM... |
|---|---|---|---|
| "rt rimamuhamad cds mms dioxidodecloro medicina barata curar corto tiempo covid19" | medicina[] | 2 | ? |
| "rt rimamuhamad cds mms dioxidodecloro medicina barata curar corto tiempo covid19" | barata[] | 2 | ? |
| "rt rimamuhamad cds mms dioxidodecloro medicina barata curar corto tiempo covid19" | curar[POSIT... | 2 | POSITIVE |
| "rt rimamuhamad cds mms dioxidodecloro medicina barata curar corto tiempo covid19" | corto[NEGA... | 2 | NEGATIVE |
| "rt rimamuhamad cds mms dioxidodecloro medicina barata curar corto tiempo covid19" | tiempo[NEG... | 2 | NEGATIVE |
| "rt rimamuhamad cds mms dioxidodecloro medicina barata curar corto tiempo covid19" | covid19[] | 2 | ? |
| "cfcuartero01 ánimo fuerza te rindas informate mms cds" | cfcuartero01[] | 2 | ? |
| "cfcuartero01 ánimo fuerza te rindas informate mms cds" | ánimo[] | 2 | ? |

*Figura 5: Taula amb els termes, el document preprocessat al qual pertanyen, la seva freqüència i l'etiqueta de sentiment*

Tenint la taula de freqüències, també podem crear un núvol de paraules o “tag cloud”. Es tracta d’una representació gràfica de les paraules més utilitzades en un conjunt de dades de text. Habitualment es representa com una imatge que conté les paraules més utilitzades, amb una mida proporcional a la seva freqüència.


*Figura 6: Núvol de paraules dels document preprocessats*



UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Escola d'Enginyeria de Barcelona Est

Seguidament, s'agrupen els missatges, es conta el nombre de paraules positives i negatives que conté cada un i a partir d'això se li assigna una etiqueta de sentiment al missatge, segons si té més paraules d'una o de l'altra categoria.

Per fer-ho, utilitzarem dos nodes diferents; un d'ells s'anomena *Rule Engine* i permet crear un seguit de condicions i definir el valor que ha de tenir el resultat si aquesta es compleix. Per a totes les files que s'ajustin a la condició, se'ls afegeix el resultat en una nova columna.

L'altre node és el *Group By*, que agrupa les files que tenen el mateix valor en una columna determinada. Per a la resta de columnes, es pot escollir entre mantenir-les, eliminar-les o agrupar-les d'alguna manera, per exemple, sumant tots els seus valors o calculant la mitjana aritmètica, entre d'altres.

Els passos que seguim per assignar les categories als missatges són:

- Amb el node *Rule Engine*, es crea una columna anomenada "Negatives" que conté un 1 si el sentiment és negatiu i un 0 en qualsevol altre cas. Seguim exactament el mateix procediment amb els positius, creant una columna anomenada "Positives".
- Ara, amb el node *Group By,* s'agrupen en una sola fila les dades que coincideixen en missatge i en autor (que les columnes "Document" i "user/screen name" siguin iguals), i per a cada un sumem els nombres de les columnes "Negatives", "Positives" i "TF abs".
- I amb un altre node del tipus *Rule Engine*, crearem una nova columna anomenada "Sentiment", que tindrà el valor de "Positive" si la columna de Sum(Positives) és major que la de Sum(Negatives) i el valor de "Negative" en cas contrari.

| Preprocessed Document | Sum(TF... | Sum(Ne... | Sum(Po... | Sentiment |
|---|---|---|---|---|
| "victoria luzdioxidodecloroensayosclinicos mms cds nadapersonal vivianacanosa dioxidodecloro..." | 16 | 0 | 1 | Positive |
| "rt yaztorq sentadav fueguinotero excelentevivo argentinafirmar pasame firmoelaboro mmse..." | 26 | 0 | 2 | Positive |
| "sentadav fueguinotero excelentevivo argentinafirmar pasame firmoelaboro mm..." | 20 | 0 | 2 | Positive |
| "mariloart gaviotaomnipot1 pcomentator dióxido cloro cds clorito sodio mms" | 18 | 0 | 0 | ? |
| "rt pdrolivermanas andreas kalcker divulgación cdsmms 8 oct 1138" | 18 | 0 | 0 | ? |
| "amonterodel saludsinbulos micromunidad farmaenfurecida ojala hubiese revistas gracias revi..." | 18 | 0 | 1 | Positive |
| "julilippi98 tito9k9 abifuentes0 tambonic turbio formaron religión cdsmmsdióxido..." | 16 | 1 | 0 | Negative |
| "mms evidencias cure covid-19 supuestos testimonios s..." | 14 | 0 | 1 | Positive |
| "rt matimatarredona anticlo2 hijos toman dióxido cloro crisis curativas quitan parásitos..." | 22 | 1 | 0 | Negative |
| "rt sandra12102017 dioxidodecloro cds mms geoestrategia cdshealscovid19 esperanzaimportant... | 26 | 0 | 1 | Positive |
| "rt sandra12102017 dioxidodecloro cds mms geoestrategia cdshealscovid19 esperanzaimportant... | 26 | 0 | 1 | Positive |
| "dioxidodecloro cds mms geoestrategia cdshealscovid19 esperanzaimportante grupo médicos mal... | 22 | 0 | 1 | Positive |
| "rt diegonunezok discriminan indios bolivianos inteligentes juntos" | 14 | 0 | 0 | ? |
| "rt enricbernedaben 153 casos efectos adversos 30 tipos mmsdióxido cloro obser... | 34 | 0 | 0 | ? |
| "randomdragon enricbernedaben salvagarca4 gemmspp carmenvivomivid andreaskaicker ipodmu... | 16 | 0 | 0 | ? |
| "rt enricbernedaben 153 casos efectos adversos 30 tipos mmsdióxido cloro obser... | 34 | 0 | 0 | ? |
| "rt enricbernedaben 153 casos efectos adversos 30 tipos mmsdióxido cloro obser... | 34 | 0 | 0 | ? |
| "miguelmcminn combalears salutgoib goib sanidadgob tema proselitismo malo me..." | 18 | 1 | 0 | Negative |
| "rt matimatarredona saliendo voces alertan eslejía dióxido clorommscds panacea medicamentosa ... | 22 | 1 | 0 | Negative |
| "saliendo voces alertan eslejía dióxido clorommscds panacea medicamentosa es..." | 18 | 1 | 0 | Negative |
| "twitter claro ununa jf mestre mal periodistaescrito proselitismo voraz eslejía e..." | 24 | 2 | 1 | Negative |
| "glezanna1993 rotoledoc eliminado tuit móvil horrible twitter quise dec..." | 18 | 1 | 0 | Negative |
| "rt matimatarredona twitter lunesdelejía insistiendo eslejía dióxido cloroclo2mmscds" | 16 | 1 | 0 | Negative |
| "twitter lunesdelejía insistiendo eslejía dióxido cloroclo2mmscds" | 12 | 1 | 0 | Negative |
| "rt enricbernedaben 153 casos efectos adversos 30 tipos mmsdióxido cloro obser... | 34 | 0 | 0 | ? |
| "rt gamezquitac ¿cómo produce dióxido cloro verdad llega sangre oxigena produce cloritos oxida ... | 28 | 0 | 1 | Positive |
| "rt maerquintero dioxidodecloromms solución mineral milagrosa" | 14 | 0 | 0 | ? |

### 4.4.3. Modificació dels diccionaris

La modificació de les paraules que incloem en els diccionaris es decideixen a partir de la precisió amb que aquests classifiquen les dades. Els passos a seguir per afegir o treure paraules són els següents:

1. S'etiqueten les paraules amb els diccionaris actuals.
2. S'agafa una mostra de 100 missatges de cada una de les tres categories: positiu, negatiu i sense classificar.
3. A partir de les mostres anteriors, s'avalua la precisió de la classificació. Per als missatges mal classificats, s'escull una o més paraules clau que estigui mal etiquetada, o bé que no ho estigui.
4. A continuació es filtren tots els missatges que contenen la paraula en qüestió. Si la majoria (>70%) corresponen a un sentiment determinat (positiu o negatiu), s'afegeix aquesta paraula al diccionari en qüestió. Si no, es manté la paraula tal i com estava.

L'objectiu d'aquest procediment és disminuir el percentatge de tweets mal classificats, així com evitar que hi hagi missatges sense classificar.

Un exemple: el terme "eslejía" inicialment no estava assignat a cap sentiment. Després de detectar que alguns missatges negatius que no estaven classificats contenien aquest terme, es va filtrar tots els missatges que el contenien. En total, de 415 tweets n'hi havia 2 de positius i 413 de negatius (99,5%). Per aquest motiu, es va afegir el terme al diccionari negatiu, amb el node *Table Creator* (veure la Figura 3).

## 4.5. Anàlisi de sentiment: Algoritmes de classificació

En aquesta part s'utilitzaran les dades amb les etiquetes assignades a l'apartat anterior per aplicar algoritmes d'aprenentatge supervisat que prediguin a quina categoria pertany cada tweet. Els algoritmes s'apliquen en dues fases: primer s'agafa un conjunt de les dades que s'utilitza per crear el model. Posteriorment, amb el model creat, es prediu la categoria de les dades restants. Com que aquestes ja tenen una categoria assignada, podrem comparar el valor d'aquesta amb el valor de la predicció per saber quina és la precisió del model.

La sortida de l'algoritme serà una variable categòrica que tindrà dos possibles valors: positiu o negatiu, per aquest motiu necessitem un algoritme de classificació. La majoria d'aquests algoritmes operen amb variables estructurades, mentre que la variable de la que disposem en aquest moment (el text), és no estructurada. Per això haurem tornar a processar les dades, perquè s'adeqüin al format requerit.

### 4.5.1. Preprocessament

En aquest punt, les dades de les que disposem són el document preprocessat i l'etiqueta de sentiment. El primer, que és la variable d'entrada a l'algoritme, és de tipus no estructurat, mentre que el segon és la categoria i és de tipus estructurat i qualitatiu (només té dos valors possibles). El problema és que molts algoritmes de classificació no admeten dades no estructurades, per això haurem de convertir el text (document) en una o més variables estructurades (ja siguin qualitatives o quantitatives). El que es fa és transformar els documents en un vector, de manera que cada paraula sigui una variable diferent. Per a un document concret, el valor de la variable d'una paraula serà el nombre de vegades que aquesta apareix al document.

Això es porta a terme amb el node *Document Vector*. La sortida d'aquest node és una taula amb 15875 columnes, una d'elles és el document que havíem preprocessat anteriorment (per tant és del tipus Document) i les altres són de tipus numèric Double, indiquen la freqüència amb que una paraula es troba en el document (a la pràctica, encara que siguin de tipus Double, tots els valors són enters perquè indiquen la freqüència absoluta).

| Document | - | negacionistas | coronavirus | antivacunas | consumidores | mms | cds |
|---|---|---|---|---|---|---|---|
| "- negacionistas coronavirus- antivacunas- consumidores mms cds- detractores 5g¿qué pasan... | 8 | 2 | 2 | 2 | 2 | 2 | 2 |
| "---- mms autoregalado cumple cds d" | 0 | 0 | 0 | 0 | 0 | 2 | 2 |
| "-la propuesta metió secta genesis ii promotora cds mms curas milagrosas curas falsas covid..." | 0 | 0 | 0 | 0 | 0 | 2 | 2 |
| "-otro medicamento costo povidone-iodine 125 detiene c-19- cds antiguo mms opción..." | 2 | 0 | 0 | 0 | 0 | 2 | 2 |
| "0diestro2 janis19022 dianasibilla diondioniss llcmoneda cds mas tragableyo usado mms a..." | 0 | 0 | 0 | 0 | 0 | 2 | 2 |
| "0diestro2 janis19022 dianasibilla diondioniss llcmoneda escds tb usar mms" | 0 | 0 | 0 | 0 | 0 | 2 | 0 |
| "0justanotherone andreas kalcker años promocionando mms cds curar pregunto qué..." | 0 | 0 | 0 | 0 | 0 | 2 | 2 |
| "0justanotherone deberían mundo debieran testimonio trabajos arbitrados válidos c..." | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| "0laseria hermannpaz loudefrogg elhuffpost eslejía dióxido clorommscds●sclo2 lejía tiene..." | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| "1 ●urgente publicaciones científicas médicas dióxidodecloro cds mms ¿xq interés censu..." | 0 | 0 | 0 | 0 | 0 | 2 | 2 |
| "1 ●urgente publicaciones científicas médicas dióxidodecloro cds mms ¿xq interés censu..." | 0 | 0 | 0 | 0 | 0 | 2 | 2 |
| "100cia duda tratamiento standard hipopotasemia guías clínicas plátano lu..." | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| "100cia evento prolongación eventos empezaron julio forcades muy..." | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| "100cia foros pongo eslejía clo2 dióxido clorommscds●s ot..." | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| "100cia glofc vaicondios apetp ferfrias cescept jvicenteprieto ejmolinac homeopatíaosteopatí... | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| "100cia noplandemia pablomarbar82 fucoxanti errekara eleperdido sabes muuuuucho sabes ... | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| "100cia pdf montón "protocolos" elaboración "tratamiento enfermedades"..." | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| "100cia revista preocupante artículos enfocados promocionarproselitismo de..." | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| "100cia salvagarca4 carmenvivomivid sachastn randomdragon gemmspp ngelbla79705027 ej... | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| "101tvmalaga utilizaran mms cds acabaría problema solución enfrente ojos quieren..." | 0 | 0 | 0 | 0 | 0 | 2 | 2 |
| "1208966jimmy rmapalacios seguro tu apoyando vacancia congreso debía disuelto punto..." | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| "1216 mms cds milagro tenes sanar origen problema principal hacernos cargono..." | 0 | 0 | 0 | 0 | 0 | 2 | 2 |
| "124456789asd danybruno10 johnarandia resto países mundo crees farmacéuticas confabula..." | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| "124456789asd danybruno10 johnarandia vende cds mms clorito sodio 28 producir el..." | 0 | 0 | 0 | 0 | 0 | 2 | 2 |
| "137solrac grupoeldeber cuantos muertos tienes cds mms conoces hables repitas quier3nnqu... | 0 | 0 | 0 | 0 | 0 | 2 | 2 |
| "14jun2020☞☞mientras jaime delgado inicia persecución legal mediática dióxid... | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

*Figura 7: Taula amb el document vectoritzat*

El node anterior presenta el desavantatge de no mantenir totes les columnes de la taula, sinó únicament la columna del document i les noves que s'han generat. Així, en aquest procés es perd l'etiqueta de sentiment que se li havia assignat anteriorment. Per recuperar-la, el que farem serà creuar les dades de la sortida del node *Document Vector* amb les dades obtingudes després de l'assignació d'etiquetes.

Per això s'utilitza el node *Joiner*, que ajunta files de taules que tenen alguna columna igual. La taula amb els missatges etiquetats conté el document preprocessat ("Preprocessed document"), que és el mateix que està en la columna "Document" a la taula del document vectoritzat. El resultat és una taula amb 15881 columnes, que conté la informació d'ambdues taules: el nom de l'usuari, el document original, el document preprocessat, la suma total de paraules, la suma de paraules positives i negatives, l'etiqueta de sentiment i les columnes vectoritzades.

Ara, les dades s'han de filtrar per tal que continguin únicament les variables que volem estudiar amb l'algoritme. A l'apartat **4.3 Selecció de les variables** ja s'ha comentat que només s'utilitzaria el text del missatge per a realitzar l'anàlisi de sentiment. Per això filtrem la resta de columnes (node *Column Filter*, ja s'ha utilitzat abans), mantenint també la columna "Sentiment", que no es tracta d'una variable sinó de la categoria.

I filtrem també aquelles dades que no tenen cap etiqueta de sentiment. Això es realitza amb el node *Row Filter*, on s'ha d'especificar que s'han d'excloure totes aquelles files que no tinguin cap valor a la columna "Sentiment". Això ens deixa amb un total de 7384 dades (de les 16003 que es tenien inicialment).

Un cop arribats aquest punt es va provar d'executar els algoritmes de classificació, però aquests no van funcionar correctament degut que hi havia paraules que tenen símbols no reconeguts per l'algoritme

(majoritàriament emoticones). El node substituïa directament el símbol per un interrogant, la qual cosa no era un problema si la paraula contenia alguna lletra reconeguda a més d'aquest. Però si el terme consistia en un únic caràcter especial, llavors el nom de la columna passava a ser un interrogant i, com que n'hi havia uns quants que tenien aquest format, el model es trobava amb vàries columnes que tenien la mateixa capçalera.

Per solucionar el problema, donat que no hi ha una manera senzilla d'eliminar els símbols especials en KNIME, s'ha optat per endreçar les columnes alfabèticament i eliminar manualment les últimes, que són les que corresponen a les emoticones. Això es porta a terme amb dos nodes; el primer és *Column resorter*, que canvia la posició de les columnes a partir d'un criteri específic. Hem de concretar el criteri d'ordenació, que és de la A-Z. Després s'utilitza un node *Column filter* per filtrar aquestes columnes.

### 4.5.2. Validació creuada

Un cop tenim les dades preparades per processar-les amb els algoritmes de classificació, les haurem de dividir en dos conjunts. Un d'ells servirà per entrenar els models i l'altre l'utilitzarem per validar-lo.

Amb aquest mètode, existeix la possibilitat que hi hagi variabilitat en el model segons el conjunt d'entrada (el d'entrenament). Pot passar que es creï un model que estigui sobreajustat o subajustat i que per tant, funcioni bé per a les dades d'entrenament però no per a la validació.

Per solucionar aquest problema, la primera cosa que fem és que els conjunts es divideixin de manera estratificada segons el sentiment. És a dir, que al conjunt d'entrenament i al de validació hi hagi la mateixa proporció de dades etiquetades com a positives i negatives. D'aquesta manera no es pot donar el cas que al conjunt d'entrenament hi hagi una gran majoria de dades d'una o de l'altra categoria, cosa que provocaria que els missatges de la categoria minoritària no es caracteritzessin correctament.

Per altra banda, apliquem una validació creuada. Això consisteix en dividir el total del conjunt de dades en un nombre determinat de parts, i executar el model tantes vegades com conjunts. En cada iteració, s'utilitza un dels conjunts per realitzar l'entrenament i la resta per a la validació. D'aquesta manera, les mètriques que s'obtindran (com l'exactitud del model) seran més acurades i robustes, no dependran tant del conjunt que escollim per fer l'entrenament.

El tipus de validació que aplicarem s'anomena k-fold cross-validation. El nombre de conjunts en què dividirem les dades (k) serà 10.

La validació creuada és realitza amb dos nodes específics i dos més que es referiran al model (s'expliquen més endavant). El primer node específic és *X-Partitioner*, que divideix les dades en el nombre de conjunts especificat. Es pot escollir entre que els conjunts es facin de manera ordenada,

aleatòria o estratificada, com ja hem comentat triarem la última opció. La primera sortida del node serà un dels conjunts i, la segona la resta.

L'altre node que necessitem és *X-Aggregator*. El que fa aquest node és, un cop executat el model, desa la classificació resultant. Quan s'acaba d'executar, provoca que es torni a executar el node *X-Partitioner*, però amb un conjunt diferent. Així es van fent les iteracions i es van desant les dades de totes elles.

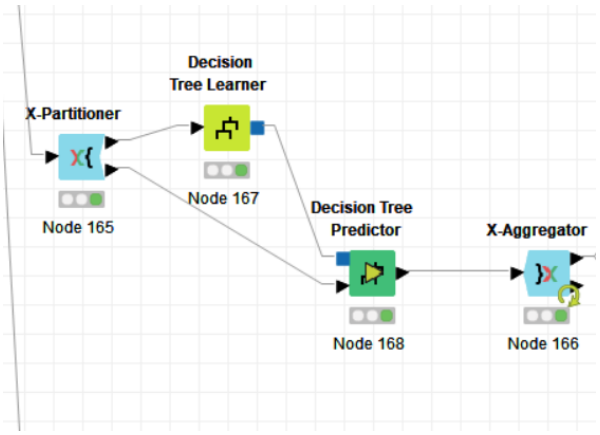


*Figura 8: Disposició dels nodes per a la validació creuada i la creació del model*

### 4.5.3. Algoritmes de classificació

Com s'ha comentat a l'apartat anterior, els algoritmes de classificació (o models) s'entrenaran i executaran de manera cíclica per portar a terme la validació creuada. En la Figura 8 podem veure que són necessaris dos nodes per realitzar aquests algoritmes; un que defineixi el model ("Learner") i un altre que, donat el model anterior l'apliqui a les dades de validació i realitzi una predicció ("Predictor").

Els dos algoritmes de classificació que s'utilitzen son *decision tree* i *random forest*, explicats posteriorment. El motiu pel qual s'han escollit és perquè, després de provar-ne varis, són els que funcionaven millor amb les dades de les que disposaven. Les principals raons per descartar els altres són la capacitat de memòria limitada (alguns algoritmes com el Support Vector Machine no es podien executar degut al gran nombre de variables)

#### 4.5.3.1. Decision tree

Un arbre de decisió o *decision tree* proporciona una descomposició jeràrquica de les dades segons una condició del valor d'una variable [13]. En el cas de la classificació de text, la condició és la presència o l'absència d'algun terme [13].

Aquest tipus d'algoritmes són bastant sensibles al conjunt d'entrada; un canvi en les dades d'entrenament pot significar que l'estructura sigui molt diferent [18]. Per això és important realitzar la validació creuada.

Els nodes que s'utilitzen són *Decision Tree Learner* i *Decision Tree Predictor*. Al primer li hem d'indicar (mitjançant la configuració) quina és la variable que volem predir ("Sentiment"). En el cas del node predictor el configurem perquè, per cada dada, mostri les probabilitats que aquesta sigui positiva o negativa. Això ho utilitzarem posteriorment per la corba ROC.

Un fragment d'un dels arbres de decisió es pot veure a continuació:

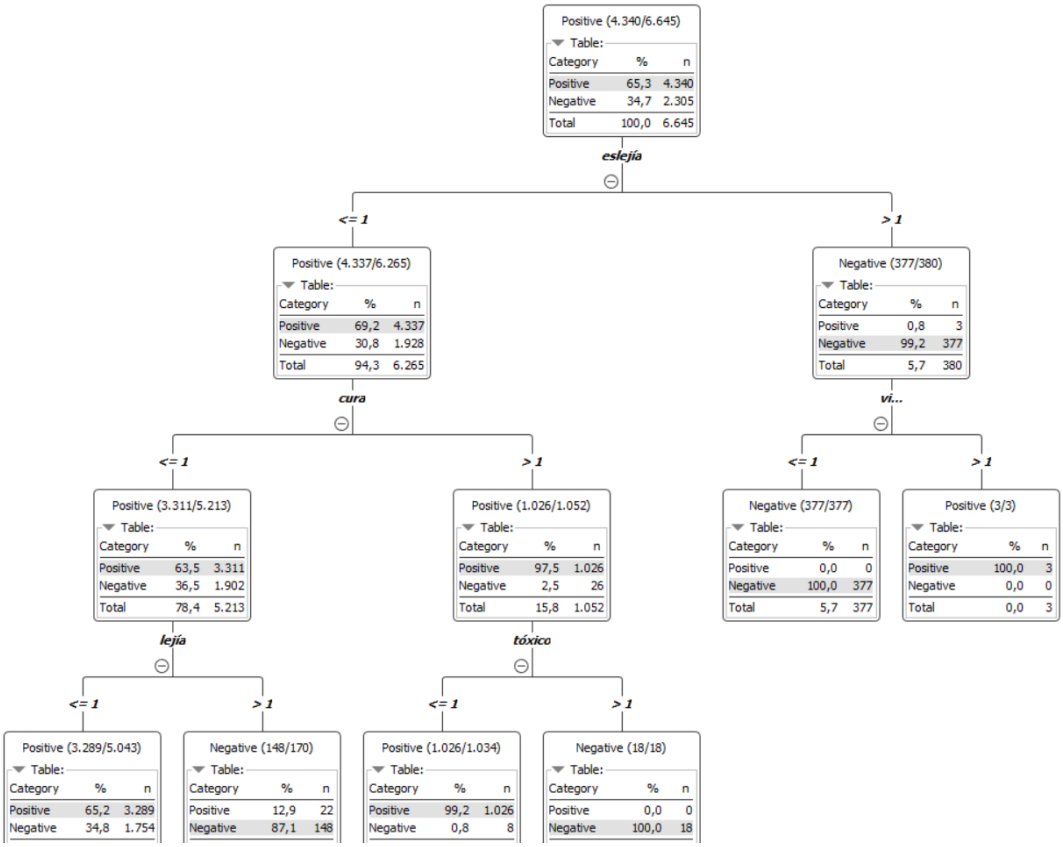


*Figura 9: Fragment del Decision Tree obtingut en una de les iteracions*

### 4.5.3.2. Random forest

L'algoritme *random forest* és molt similar al de l'arbre de decisió. De fet, el model crea diversos arbres de decisió i, per avaluar les dades de validació, els executa tots i dona com a predicció l'opció més probable [18].

Respecte a l'arbre de decisió, presenta l'avantatge que el model és molt més robust, ja que si un arbre presenta alguna desviació, la resta fan que es corregeixi [18]. Però òbviament el seu cost computacional és més elevat que realitzant un únic arbre de decisió.

A l'hora d'aplicar la validació creuada en aquest model ens trobem que el processador de l'ordinador no permet executar aquesta amb un random forest que contingui un gran nombre d'arbres (l'ordinador es queda penjat). El màxim que s'ha aconseguit fer és una classificació amb 6 arbres de decisió.

Donat que és una quantitat petita d'arbres (els random forests poden tenir-ne més de 100), s'ha executat també l'algoritme amb 100 arbres sense validació creuada. Els resultats són que l'exactitud i altres mètriques són més elevades per aquest últim (l'exactitud passa de ser 0,9515 a 0,996). Per tant, i com que es tracta d'un algoritme robust, s'ha optat per tenir un model amb més arbres de decisió i sense validació creuada.

Llavors, haurem d'afegir un altre node que divideixi les dades en dos conjunts únicament. És el node *Partitioning*, que configurem perquè faci un conjunt amb el 70% de les dades i l'altre amb les restants, de manera estratificada.

# 5. Resultats

## 5.1. Diccionaris

Amb els diccionaris inicials ja es va detectar que algunes paraules no estaven ben classificades, o bé directament no es trobaven a les llistes. Com ja s'ha comentat, els diccionaris s'han modificat per incloure més paraules o eliminar aquelles que no fossin d'utilitat. Les paraules que s'han afegit es llisten a continuació:

| **Positives** | **Negatives** |
|---|---|
| retomar | pseudo |
| tomar | pseudociencia |
| especialista | eslejía |
| oxigenar | desinformación |
| ivermectina | secta |
| terapia | prohibe |
| tratamiento | fake |
| efectivo | stoppseudociencias |
| activador | bebelejias |
| salvar | bebelejías |
| milenario | |
| plandemia | |
| ayuda | |
| mejorando | |
| yotomocds | |
| curaciones | |
| confio | |
| industriafarmacéutica | |

També s'han eliminat dues paraules del diccionari negatiu, aquestes son: *virus* y *mucho*.

El nombre total de dades classificades com a positives, negatives i dades sense classificar (SC) es pot trobar a la taula següent. Els resultats estan separats per si corresponen als diccionaris originals o als modificats.

| | Diccionaris originals | Diccionaris modificats |
|---|---|---|
| Positives | 4308 | 4835 |

| | | |
|---|---|---|
| Negatives | 2279 | 2549 |
| Sense classificar | 9416 | 8619 |

Per a cada diccionari i categoria, s'ha agafat una mostra de 100 dades i s'ha comprovat manualment quines dades estaven ben classificades i quines no, i s'ha calculat l'exactitud en aquestes mostres i les matrius de confusió. Per exemple, s'han agafat 100 mostres classificades com a negatives pels diccionaris originals, i d'aquestes s'ha detectat que 69 realment corresponen a dades amb sentiment negatiu. Però en el mateix conjunt n'hi ha 29 que són positives i que s'han classificat malament, i dues que no s'haurien d'haver classificat amb cap sentiment però que tot i així tenen una etiqueta negativa.

Encara que no són mostres representatives, i que les exactituds que calculem amb aquestes dades no es poden extrapolar per a totes les dades de la categoria, ens servirà per tenir una idea general de com de bona és la classificació que realitzen els diccionaris.

| Diccionaris originals | | | |
|---|---|---|---|
| VR / VP* | Negatiu | Positiu | SC |
| Negatiu | 69 | 47 | 43 |
| Positiu | 29 | 51 | 42 |
| SC | 2 | 2 | 15 |

| Diccionaris modificats | | | |
|---|---|---|---|
| VR / VP* | Negatiu | Positiu | SC |
| Negatiu | 76 | 38 | 47 |
| Positiu | 20 | 59 | 27 |
| SC | 4 | 3 | 26 |

*VR: Valor Real // VP: Valor Predit

Seguidament podem trobar dos exemples de missatges mal classificats i els motius de l'error:

Fals positiu (missatge classificat com a positiu però que pertany a la categoria dels negatius):

| |
|---|
| *"Es falso que se haya demostrado que el dióxido de cloro es inofensivo para el ser humano"* |

S'ha classificat com a positiu perquè conté dues paraules positives (*demostrado, inofensivo*) i una de negativa (*falso*).

Fals negatiu (missatge classificat com a negatiu però que pertany a la categoria dels positius):

| "Basta ya de mentiras 😕 No engañes más. MMS/CDS mata Covid19" |
|---|

S'ha classificat com a negatiu per les paraules *mentiras* y *mata*.

## 5.2. Algoritmes de classificació

Per l'avaluació de resultats dels algoritmes de classificació s'utilitzen com a mètriques: la corba ROC i la respectiva àrea sota la corba, l'exactitud, precisió, sensitivitat i especificitat.

La corba ROC és un tipus de gràfic que s'utilitza per conèixer el rendiment de classificacions binàries. Es representa la sensitivitat en funció de 1-especificitat, i es calcula l'àrea sota la corba. Si aquesta és propera a 1 vol dir que el classificador té un alt percentatge d'encert, en canvi, per a una àrea del 0,5 significa que l'algoritme s'assimila a un classificador aleatori.

Les altres mètriques són:

- Exactitud: És la relació entre la quantitat de dades ben classificades (verdaders positius i verdaders negatius) i les dades totals.
- Precisió: És la relació entre verdaders positius sobre tots els valors classificats com a positius [19]. Indica la probabilitat que un valor classificat com a positiu ho sigui realment.
- Sensitivitat: Relació entre els verdaders positius i tots els positius [19]. Indica quina és la probabilitat que un valor positiu sigui detectat com a tal.
- Especificitat: Relació entre els verdaders negatius i tots els negatius [19]. Indica quina és la probabilitat que un valor negatiu sigui detectat com a tal.

Com podem suposar, les mètriques anteriors tenen un valor comprès entre 0 i 1 i, seran més elevades quan millor sigui el model.

Per tal d'obtenir-les, s'utilitzen els nodes *ROC Curve* i *Scorer*, a la sortida del *X-Aggregator* (quan acaba el bucle de la validació creuada). Per a *Scorer*, la configuració consisteix en assignar quina és la variable inicial i quina la predicció. A partir d'aquestes dades, calcula quantes dades estan classificades correctament, quants son falsos negatius i quants falsos positius.

Per fer la corba ROC no es necessària la predicció, sinó les probabilitats de que cada una de les dades sigui positiva. Ja hem afegit una columna d'aquestes característiques al predictor, així ja es pot configurar el node.

| *Mètrica / Algoritme* | *Decision Tree* | *Random Forest* |
|---|---|---|
| *Exactitud* | 0,9217 | 0,9964 |
| *Precisió* | 0,9313 | 0,9959 |
| *Sensitivitat* | 0,9505 | 0,9986 |
| *Especificitat* | 0,8670 | 0,9922 |
| *Àrea sota la corba ROC (AUC)* | 0,9585 | 0,9999 |

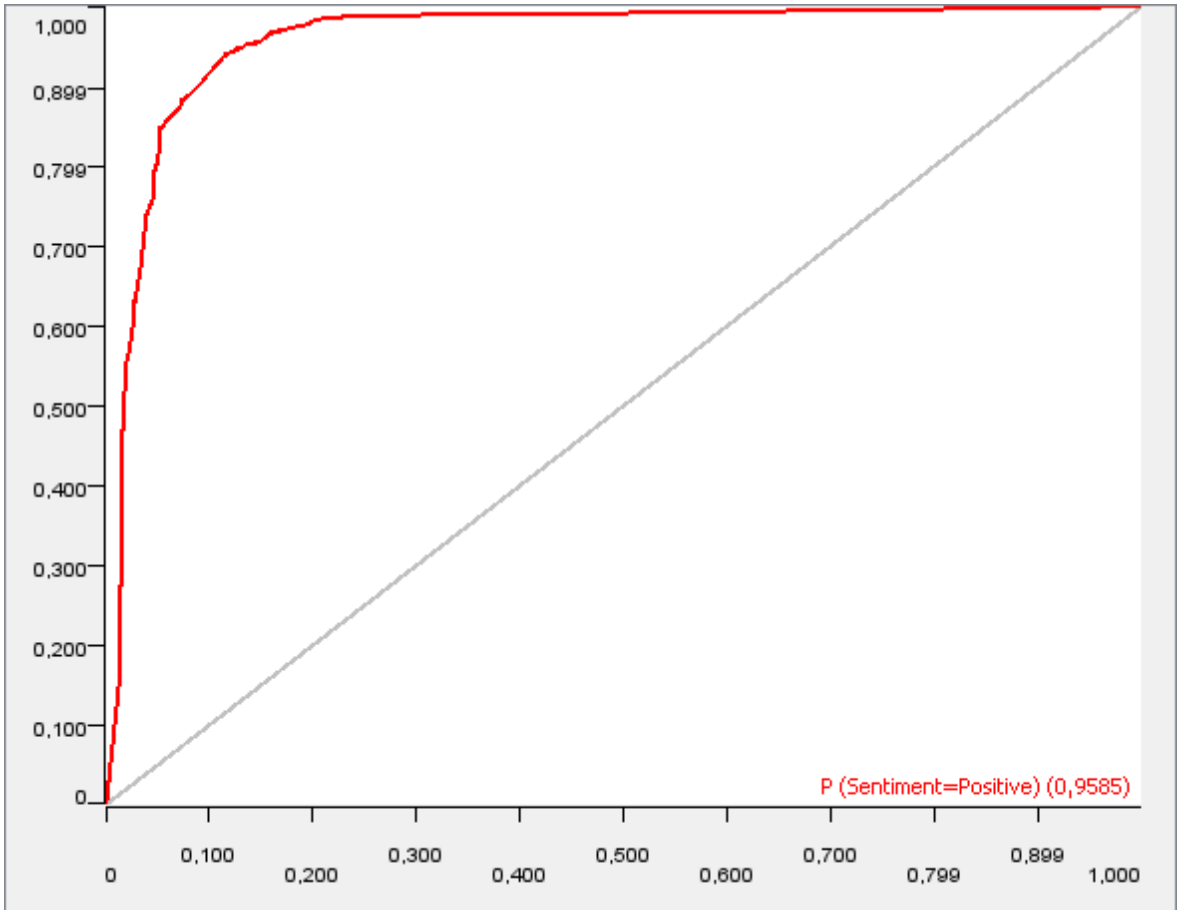


*Figura 10: Corba ROC del classificador Decision Tree*

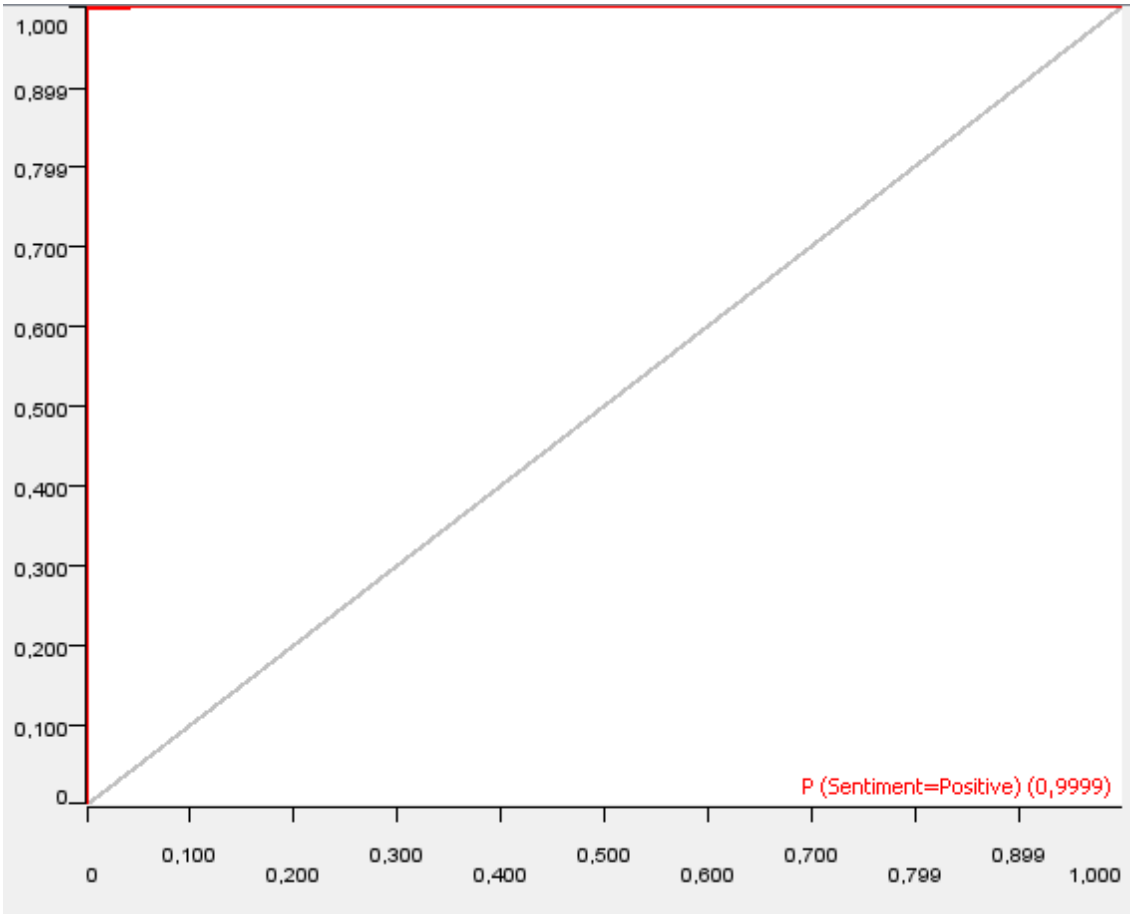


*Figura 11: Corba ROC del classificador Random Forest*

## 5.3. Discussió

Respecte a l'anàlisi de sentiment amb diccionaris, el primer que sobta és que de la mostra de dades sense classificar, a la majoria sí que se'ls podria assignar una categoria (un 85% en el cas dels diccionaris originals i un 74% amb els diccionaris modificats).

La idea era que només es quedessin sense classificar aquells missatges que, amb el text que contenien, no fos possible determinar-ne un sentiment. En són exemples els missatges molt curts, o bé aquells que únicament inclouen mencions a persones o a termes, però sense formar una frase amb sentit. Un exemple de tweet que no està classificat:

| "@lorosoria CDS MMS Dioxido de Cloro –" |
|---|

Els missatges que es queden sense classificar s'haurien de reduir al màxim, ja que són dades que s'han adquirit però que s'acaben perdent, moltes vegades a causa de no disposar d'un diccionari adequat. En aquest sentit veiem una millora significativa quan apliquem els diccionaris modificats. Les dades que

no tenen cap categoria associada passen de ser 9416 a 8619, és a dir, s'han classificat gairebé 800 dades modificant 30 paraules dels diccionaris.

Respecte a les dades classificades com a negatives o positives, el percentatge d'encert de la mostra augmenta quan es modifiquen els diccionaris amb noves paraules, segons podem observar a la matriu de confusió.

En relació als algoritmes de classificació, les mètriques mostren uns molts bons resultats de classificació, millors en el Random Forest que en el Decision Tree, però per sobre del 90% d'encert general. La més baixa és l'especificitat en el Decision Tree, que està per sota de la resta de resultats. El que això significa és que hi ha diversos valors negatius que s'estan classificant com a positius, tot i així la taxa d'encert és bastant alta.

Ara bé, s'ha d'anar amb compte amb la interpretació que es fa d'aquests resultats. Recordem que les dades etiquetades provenien de l'anterior anàlisi amb diccionaris i que, per tant, no representen en tots els casos la categoria real a la que hauria de pertànyer el missatge. És a dir, podria ser que un missatge estigués originalment etiquetat com a negatiu, la predicció de l'algoritme fos negatiu també, però que en realitat el seu sentiment fos positiu.

També és interessant notar que s'han eliminat prèviament aquelles dades que no tenien categoria, que eren més de la meitat. Per això és possible que es produeixi una desviació, ja que les dades utilitzades corresponen a aquelles que tenen com a mínim una paraula de les que hi ha als diccionaris. Existeix la possibilitat que les dades que s'hagin utilitzat com a entrada dels algoritmes de classificació siguin menys complexes de classificar que les restants.

Per tant, no podem afirmar que els algoritmes de classificació facin una bona predicció de les categories reals de les dades, si no que fan una bona predicció de les categories atorgades pels diccionaris. S'hauria de comprovar si, etiquetant les dades manualment se seguirien obtenint prediccions fiables.

## 5.4. Propostes de millora

La principal proposta de millora del treball és l'ampliació dels diccionaris. Els que teníem inicialment, tot i comptar amb moltes paraules, no eren específics per al tema que estàvem treballant, per això el rendiment no era gaire elevat. A més, hi havia moltes dades sense classificar. S'ha comprovat com afegint unes poques paraules al diccionari que siguin específiques del conjunt de dades es pot millorar el rendiment de l'algoritme i alhora augmentar el nombre de dades classificades. Ara bé, aquest procés s'ha fet manualment i és bastant lent, per això seria interessant poder fer-ho de manera automatitzada.

Una altra cosa que seria interessant de cara al preprocessament de les dades seria poder extreure l'arrel de les paraules. Treballar amb les paraules completes provoca que els algoritmes detectin com a diferents dues paraules que són de la mateixa família i que per tant tenen el mateix significat. En són exemples les paraules en plural o les diferents conjugacions dels verbs. KNIME disposa d'un node que ho porta a terme, però només hi ha la versió en anglès.

També en relació al preprocessament, al final de l'anàlisi s'ha detectat que hi havia paraules que no s'havien separat entre sí en els missatges, i que figuraven com una sola. El problema ve que es va especificar al 'word tokenizer' (al node *Strings to Document*) que separés les paraules per els espais en blanc, quan n'hi ha algunes que estan separades per tabuladors.

# 6. Anàlisi de l'impacte ambiental

El present treball és de tipus de desenvolupament de programari. No s'han realitzat cap procés de fabricació, elaboració o transformació de materials. No s'han generat residus, ni emissions al medi ambient, i podem dir que el seu impacte es redueix al que pugui generar l'ordinador quan sigui substituït i al consum d'energia elèctrica durant la realització del TFG.

Personalment, l'únic que s'ha utilitzat és un ordinador portàtil, que necessita ser endollat aproximadament cada quatre hores, i que es carrega completament en una hora. Durant les hores de llum, aquesta energia prové en gran part d'energies renovables (energia solar), però durant la nit només un 3,7% de l'energia que s'utilitza és d'aquest tipus.

Per descomptat, per la realització del treball també ha estat imprescindible altre hardware extern i sobre el qual no es té un control; en són exemples els servidors de les eines que s'han utilitzat (KNIME, Twitter i Python). Però és pràcticament impossible estimar l'impacte que ha tingut el desenvolupament de l'aplicació en aquests servidors.

# Conclusions

Una de les primeres conclusions a les que s'ha arribat amb la realització d'aquest treball és que les dades no estructurades presenten una complexitat important a l'hora de processar-se. Un preprocessament general deixa algunes dades en un format incorrecte (com per exemple, les dades separades per un tabulador en comptes d'un espai) i aquestes poden provocar una desviació a l'hora d'elaborar els models o bé que la dada en qüestió no es pugui processar (com ens ha passat). Per tant, és un tema en el que s'hauria d'aprofundir bastant abans de realitzar qualsevol tipus d'anàlisi, conèixer bé les característiques que presenten totes i cada una de les dades, i provar de descartar a l'inici aquelles que presenten variacions.

Una de les conclusions que es pot extreure de l'anàlisi de sentiment basat en el lèxic és que, perquè classifiqui correctament les dades, cal uns diccionaris adaptats a la temàtica que s'està estudiant. Mentre que els diccionaris genèrics estan bé com a punt de partida, afegir termes nous o modificar-los pot aportar una diferència qualitativa en les classificacions que es realitzen. Un avantatge d'aquest algoritme és que es pot aplicar de manera individual, així que, amb uns diccionaris més complexos podria ser una bona eina de detecció per al sentiment de missatges.

I per a l'anàlisi de sentiment basat en algoritmes d'aprenentatge automàtic, amb Decision Tree i Random Forest, podem concloure que la seva exactitud com a classificador és molt bona en els dos que s'han utilitzat. La taxa d'error és molt baixa i amb la validació creuada s'ha pogut comprovar com els models no estaven sobreajustats.

No obstant això, aquests algoritmes necessiten que les dades estiguin prèviament etiquetades, per la qual cosa no són adequats per ser utilitzats de manera exclusiva. Les dades de Twitter, i en general les de qualsevol xarxa social no acostumen a venir etiquetades amb una de les categories que es voldran predir. És per això que una possible aplicació d'aquests algoritmes seria, per exemple, la de classificar missatges nous que es van publicant un cop ja tenim una gran base de dades classificada.

Una altra eina interessant de cara a comprendre els missatges que es publiquen sobre un tema és el núvol de paraules. Ja que, a més de conèixer el sentiment de manera general, saber quins són els termes més utilitzats i a quin sentiment s'associen pot aportar una comprensió més profunda de les dades sense haver-les de llegir totes.

Finalment, podem dir que s'han complert els objectius inicials en quant a la implementació i avaluació dels algoritmes utilitzats per a un cas pràctic, i que l'aplicació desenvolupada, amb les millores que s'han comentat podria servir com a classificador de sentiment per a missatges relacionats amb MMS.

# Pressupost

En general, hi ha pocs conceptes amb cost associats al desenvolupament de l'aplicació. D'una banda, hi ha les hores dedicades per part de l'estudiant (únicament s'han contat aquelles en què s'han obtingut les dades i s'ha desenvolupat l'aplicació). Per altra banda, s'ha calculat la devaluació de l'ordinador portàtil utilitzat. S'ha realitzat un càlcul estimatiu amb el seu preu de venta i el nombre de mesos que s'ha utilitzat per l'aplicació (4).

Tant KNIME com Python són de software lliure (gratuïts), mentre que per obtenir les dades de Twitter que es necessitaven sí que es va pagar una subscripció.

| **Concepte** | **Capítol** | **Preu unitari (€)** | **Unitats** | **Preu (€)** |
|---|---|---|---|---|
| **Hores de desenvolupament de l'aplicació** | Metodologia | 15 | 350 | 5250 |
| **Devaluació ordinador portàtil per al desenvolupament de l'aplicació** | Metodologia | 40 | 4 | 160 |
| **Un mes de subscripció al compte Premium de Twitter API** | Obtenció de dades | 91,15 | 1 | 91,15 |
| **Total** | | | | 5501,15 |

# Referències i bibliografia

[1] BIN NAEEM, S., BHATTI, R., KHAN, A (2020). *An exploration of how fake news is taking over social media and putting public health at risk*. A: *Health Information and Libraries Journal*. Disponible a: https://doi.org/10.1111/hir.12320 [Data d'última consulta: 16 de juny de 2021]

[2] SALUD SIN BULOS, DOCTORALIA (2019). *II Estudio sobre Bulos en Salud*. Disponible a: https://saludsinbulos.com/wp-content/uploads/2019/11/es-II-estudio-bulos-salud.pdf [Data d'última consulta: 16 de juny de 2021]

[3] LARDIERI, A., CHENG, C., JONES, S.C., MCCULLEY, L. (2020). *Harmful effects of chlorine dioxide exposure,* A: *Clinical Toxicology*. Disponible a: https://doi.org/10.1080/15563650.2020.1818767 [Data d'última consulta: 16 de juny de 2021]

[4] Home. *Jim Humble*. Disponible a: https://jimhumble.co/ [Data d'última consulta: 16 de juny de 2021]

[5] *Determination of the Effectiveness of Oral Chlorine Dioxide in the Treatment of COVID 19.* Clinicaltrials.gov. Disponible a: https://clinicaltrials.gov/ct2/show/NCT04343742?term=chlorine+dioxide&draw=8&rank=1 [Data d'última consulta: 16 de juny de 2021]

[6] *An Outpatient Study Investigating Non-prescription Treatments for COVID-19 (PROFACT-01)*. Clinicaltrials.gov. Disponible a: https://clinicaltrials.gov/ct2/show/NCT04621149?term=chlorine+dioxide&draw=3&rank=12 [Data d'última consulta: 16 de juny de 2021]

[7] INSIGNARES-CARRIONE, E. et. al. (2021). *Determination of the Effectiveness of Chlorine Dioxide in the Treatment of COVID 19.* A: J*ournal of Molecular and Genetic Medicine.* Disponible a: https://www.hilarispublisher.com/open-access/determination-of-the-effectiveness-of-chlorine-dioxide-in-the-treatment-of-covid-19.pdf [Data d'última consulta: 16 de juny de 2021]

[8] *Alerta de medicamentos ilegales, Nº 05/10 – MMS (Miracle Mineral Solution).* Agencia española de medicamentos y productos sanitarios. Disponible a: https://www.aemps.gob.es/informa/notasinformativas/medicamentosusohumano-3/medIlegales/2010/ni_muh_ilegales_05-2010/?lang=ca [Data d'última consulta: 16 de juny de 2021]

[9] *Danger: Don't Drink Miracle Mineral Solution or Similar Products.* U.S. Food & Drug Administration. Disponible a: https://www.fda.gov/consumers/consumer-updates/danger-dont-drink-miracle-mineral-solution-or-similar-products [Data d'última consulta: 16 de juny de 2021]

[10] OSMAN, M. (2021). *Estadísticas Impresionantes de Twitter y Datos Importantes Sobre Nuestra Red Favorita.* Disponible a: https://kinsta.com/es/blog/estadisticas-twitter/ [Data d'última consulta: 16 de juny de 2021]

[11] *More about restricted uses of the Twitter APIs*. Twitter Developer Platform. Disponible a: https://developer.twitter.com/en/developer-terms/more-on-restricted-use-cases [Data d'última consulta: 16 de juny de 2021]

[12] POZZI, F.A., FERSINI, E., MESSINA, E., LIU, B. (2017). *Sentiment Analysis in Social Networks.* Elsevier.

[15] MEDHAT, W., HASSAN, A., KORASHY, H. (2014). *Sentiment analysis algorithms and applications: A survey.* A: *Ain Shams Engineering Journal*. Disponible a: https://doi.org/10.1016/j.asej.2014.04.011 [Data d'última consulta: 16 de juny de 2021]

[14] PORTER, S. J. (2019). *KNIME Analytics Platform is the "killer app" for machine learning and statistics*. Towards Data Science. Disponible a: https://towardsdatascience.com/knime-desktop-the-killer-app-for-machine-learning-cb07dbef1375 [Data d'última consulta: 16 de juny de 2021]

[15] RASTOGI, P. (2020). *Extracting Tweets Using Twitter Premium Search API and Python*. Medium. Disponible a: https://medium.com/swlh/extracting-tweets-using-twitter-premium-search-api-and-python-2d025144e8a4 [Data d'última consulta: 16 de juny de 2021]

[16] GARCÍA, S., LUENGO, J., HERRERA, F. (2015). *Data Preprocessing in Data Mining.* Springer: Intelligent Systems Reference Library. Vol. 72.

[17] *Lexicon Based Approach for Sentiment Analysis*. KNIME Hub. Disponible a: https://hub.knime.com/knime/spaces/Examples/latest/08_Other_Analytics_Types/01_Text_Processing/26_Sentiment_Analysis_Lexicon_Based_Approach~zp_hhUROHNXToZHX [Data d'última consulta: 16 de juny de 2021]

[18] YIU, T. (2019). *Understanding Random Forest.* Towards Data Science. Disponible a: https://towardsdatascience.com/understanding-random-forest-58381e0602d2 [Data d'última consulta: 16 de juny de 2021]

[19] RODELLAR, J., ACEVEDO, A. (2021). *Clasificador Bayes y Análisis Discriminante Lineal.* Aprendizaje Bioestadístico.

[20] *Python - Word Tokenization*. Tutorials Point. Disponible a: https://www.tutorialspoint.com/python_data_science/python_word_tokenization.htm [Data d'última consulta: 16 de juny de 2021]

ÁVILA RODRÍGUEZ, M. P. (2020) *Análisis de tweets y su influencia en los seguros de vida en el ámbito colombiano.* Universitat Politècnica de València. TFM.

RAJPUT, Q., HAIDER, S., GHANI, S. (2016). *Lexicon-Based Sentiment Analysis of Teachers' Evaluation.* A: *Applied Computational Intelligence and Soft Computing*. Disponible a: https://doi.org/10.1155/2016/2385429 [Data d'última consulta: 16 de juny de 2021]

MIHANOVIĆ, A., GABELICA, H., KRSTIĆ, Z. (2014). *Big Data and Sentiment Analysis using KNIME: Online Reviews vs. Social Media*. A: *International Convention on Information and Communication Technology, Electronics and Microelectronics*. Disponible a: https://ieeexplore.ieee.org/abstract/document/6859797 [Data d'última consulta: 16 de juny de 2021]

HOFMANN, M., CHISHOLM, A. (2016). *Text Mining and Visualization Case Studies Using Open-Source Tools.* CRC Press Taylor & Francis.

# Annex A: Codi de l'extracció de dades

```
pip install searchtweets

import yaml
config = dict(
   search_tweets_api = dict(
      account_type = 'premium',
      endpoint = 'https://api.twitter.com/1.1/tweets/search/fullarchive/research.json',
      consumer_key = 'J3H5F40mQmGx8E2AJRhVJMhgO',
      consumer_secret = 'DrWi1ylH5NaIDxBph7Zzrwz3PHAqBk7CIFjRltgSxVHi8yHbb3'
   )
)
with open('twitter_keys_fullarchive.yaml', 'w') as config_file:
   yaml.dump(config, config_file, default_flow_style=False)


from searchtweets import load_credentials

premium_search_args = load_credentials("twitter_keys_fullarchive.yaml",
                           yaml_key="search_tweets_api",
                           env_overwrite=False)
print(premium_search_args)

from searchtweets import gen_rule_payload
query = ("mms (cds OR dioxidodecloro OR chlorinedioxide)")
rule = gen_rule_payload(query, results_per_call=500, from_date="2020-08-01", to_date="2020-08-
13")

from searchtweets import ResultStream

rs = ResultStream(rule_payload=rule,
            max_results=1000,
            **premium_search_args)

import json
with open('tweetsData30.json', 'a', encoding='utf-8') as f:
   for tweet in rs.stream():
```

```
        json.dump(tweet, f)
        f.write('\n')
print('done')
```